AFRL-RH-WP-TR-2016-0043

IMPACT OF INDIVIDUAL DIFFERENCES ON RELIANCE OPTIMIZATION

Gloria Calhoun, Gregory Funke
Air Force Research Laboratory

Gerald Matthews, Ryan Wohleber, Jinchao Lin
Institute for Simulation & Training, University of Central Florida

Chung-yiu Peter Chiu
University of Cincinnati

Heath Ruff
Infoscitex

Interim Report

May 2016

AIR FORCE RESEARCH LABORATORY
711 HUMAN PERFORMANCE WING,
AIRMAN SYSTEMS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

Notice and Signature Page

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2016-0043 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//signed//
GLORIA L. CALHOUN
Program Manager
Supervisory Control and Cognition Branch

//signed//
JASON B. CLARK
Chief, Supervisory Control and Cognition Branch
Warfighter Interface Division

//signed//
WILLIAM E. RUSSELL
Chief, Warfighter Interface Division
Human Effectiveness Directorate

This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. REPORT DATE ***(DD-MM-YY)***	**2. REPORT TYPE**	**3. DATES COVERED** ***(From – To)***
31-05-16	Interim	From 2-03-2012 to 31-05-2016

4. TITLE AND SUBTITLE
IMPACT OF INDIVIDUAL DIFFERENCES ON RELIANCE OPTIMIZATION

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Gloria Calhoun, Gregory Funke (Air Force Research Laboratory)
Gerald Matthews, Ryan Wohleber, Jinchao Lin (Institute for Simulation & Training, University of Central Florida)
Chung-yiu, Peter Chiu (University of Cincinnati)
Heath Ruff (Infoscitex)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
H0JC (53290904)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
711 HPW/RHCI
2210 Eighth Street
Wright-Patterson AFB OH 45433-7511

8. PERFORMING ORGANIZATION REPORT NUMBER
AFRL-RH-WP-TR-2016-0043

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
711 HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AFB OH 45433

10. SPONSORING/MONITORING AGENCY ACRONYM(S)
711 HPW/RHCI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES
This is an Interim Report for Workunit 53290904 and the Final Report for AFOSR LRIR A9550-13-1-0016. A portion of this research was performed by the University of Cincinnati and the University of Central Florida via FA9550-13-1-0016.
88ABW Cleared 07/08/2016; 88ABW-2016-3397.

14. ABSTRACT (Maximum 200 words)
Envisioned Air Force operations will involve single operator supervision of multiple remotely piloted aircraft (RPAs). However, the ability of the operator to benefit from the automation's support can be jeopardized by inappropriate (over- and under-) reliance on automation. In Air Force Office of Scientific Research (AFOSR) linked funded efforts, the Air Force Research Laboratory developed methodologies and configurations of an RPA simulation to support experimental studies that were collected at partner academic laboratories. The objective of these linked efforts was to explore human performance and operator reliance on automation with various levels and reliabilities of automation, across several states of fatigue. The results provide a better understanding of the circumstances under which individual differences (e.g., video game experience), fatigue and automation characteristics may interact to produce inappropriate reliance on automation. In addition, the utility of eye tracking to diagnose suboptimal use of automation was explored.

15. SUBJECT TERMS
Remotely Piloted Aircraft, Fatigue, Automation Reliance, Level of Automation, Eye Tracking, Adaptive Interface

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Gloria Calhoun
Unclassified	Unclassified	Unclassified	SAR	48	**19b. TELEPHONE NUMBER** *(Include Area Code)*

SF Form 298 Continuation

7. PERFORMING ORGANIZATION(S) AND ADDRESS(ES)

Institute for Simulation and Training
University of Central Florida
3100 Technology Parkway
Orlando FL 32826

University of Cincinnati
2600 Clifton Avenue
Cincinnati OH 45220

Infoscitex
4027 Colonel Glenn Hwy
Beavercreek OH 45431

Table of Contents

List of Figures

List of Tables

1. ABSTRACT

Air Force operations are becoming increasingly automated, exemplified by the vision to enable single operator supervision of multiple remotely piloted aircraft (RPAs). While increased automation is a requirement for this vision, it also introduces human factors issues. In particular, the operator may become over- or under-reliant on automation. With inappropriate automation reliance, the ability of the operator to benefit from the automation's support is jeopardized, and human-automation system performance may be compromised. This Air Force Office of Scientific Research (AFOSR) funded Laboratory Research Independent Research effort was conducted in collaboration with an AFOSR grant to academia (University of Cincinnati). The Air Force effort developed methodologies and configurations of an RPA simulation to support experimental studies conducted at the University of Cincinnati and the University of Central Florida. The joint effort explored human performance and reliance on automation with various levels and reliabilities of automation across multiple states of fatigue. The role of individual difference measures in successful automation usage, including video game experience, was examined. In addition, the utility of eye tracking to diagnose fatigue and suboptimal use of automation was explored. Finally, automation configurations were evaluated to explore their effectiveness in mitigating inappropriate automation reliance. The objective of these linked efforts was to gain a better understanding of the circumstances under which individual differences, fatigue, and automation characteristics may interact to produce inappropriate reliance on automation. These results will inform methods being considered to optimize automation use that would benefit Air Force programs requiring supervisory control of autonomous systems.

2. INTRODUCTION

Air Force operations are becoming increasingly automated, exemplified by the vision to enable single operator supervision of multiple remotely piloted aircraft (RPA). Single operator control of multiple RPAs is anticipated to be a particularly time-critical, cognitively demanding multi-task work environment (Calhoun, Ruff, Draper, & Wright, 2011; Guznov, Matthews, Funke, & Dukes, 2011). In response, developments are underway to extensively automate RPA functions with the goal of enhancing the operator's ability to manage task demands. However, to realize the benefits of automation, in terms of improved mission effectiveness, an appropriate level of trust in the automation must be established and maintained (Lee & See, 2004). For instance, RPA automation carries the risk of operator over-reliance on the technology, leading to complacency effects (as shown empirically in multi-RPA simulations, e.g., Calhoun et al., 2011). However, if an operator views the automation's functioning or reliability as suspect, under-reliance may result, limiting its potential benefit, and possibly leading to a concomitant increase in operator workload. Research is needed to better understand factors influencing reliance on automation and identify candidate strategies that might optimize reliance for autonomous systems.

Reliance on automation may reflect psychological elements, such as individual differences of operators. One potentially relevant difference is the operator's experience with video games. In the most relevant studies, video game exposure was demonstrated to be positively associated with a range of sensory, perceptual, and attentional abilities that contribute to performance on tasks requiring spatial cognition (Richardson, Powers, & Bousquet, 2011; Spence & Feng, 2010). Importantly, some studies provide evidence for 'far transfer;' training on video games produced improvements on other spatial tasks that are not closely similar to the task used for training (Spence & Feng, 2010). In addition, research by Cummings, Clare, and Hart (2010) revealed that

experienced gamers might collaborate more effectively with automation. Thus, video game experience may have specific benefits for RPA operators.

Non-cognitive individual differences may also drive reliance on automation. Just as these differences have been found to influence team performance (Helmreich, Merritt, & Wilhelm, 1999), non-cognitive mediators may generalize to human-automation interactions in the supervisory control domain, with automation considered an intelligent agent teamed with human operators. Szalma and Taylor (2011) explored the five-factor model of personality in response to automation in an unmanned ground vehicle (UGV) task. Their results, in addition to those of Chen and Terrence (2009; using attentional control survey and spatial ability tests with ground robots), support incorporation of individual differences into automation design. In respect to the air domain, one multi-RPA simulation showed a strong positive correlation between a measure of participant extraversion and reliance on automation (Kidwell, Calhoun, Ruff, & Parasuraman, 2012). While the results from these and other studies suggest the importance of considering individual differences in automation design, the research also highlights the need to better understand the interplay of multiple factors (e.g., individual differences, level and reliability of automation, workload, stress, etc.). Results from Szalma and Taylor (2011) suggest that trait effects can be attenuated if participants can cross-check the automation with the "raw data" itself for performance verification. Thus, future research needs to examine how the degree to which operators have insight into the automation (transparency) moderates the relationship of traits to agreement with decision automation (see Chen et al., 2014).

In examining the interplay of factors influencing reliance on automation, one likely contributing contextual element is operator fatigue. RPA operations are often sustained over lengthy periods, during which operators may experience declining vigilance and loss of task engagement (Warm, Parasuraman, & Matthews, 2008). Field studies confirm that fatigue is a significant issue in current RPA operations (Tvaryanas & MacPherson, 2009). However, little is known about the role of fatigue in next-generation multi-RPA systems that will be more autonomous. Potentially, automation might be beneficial in supporting the fatigued pilot. Alternately, there may be a negative impact if the pilot's fatigue results in over-reliance on the automation. This may stem from fatigue encouraging passive, reactive strategies as opposed to maintaining proactive control (Hockey, Wastell, & Sauer, 1998; Matthews & Desmond, 2002). If so, fatigue may encourage excessive trust, as the operator modifies task goals to minimize personal effort.

3. RESEARCH GOALS AND OBJECTIVES

An initial research goal was to prepare a simulation of multi-RPA operation for examining effects of automation configuration and fatigue on operator use and misuse of automation. Once suitable task paradigms were developed, experimentation commenced to examine the role of individual difference factors associated with personality, trust, and video gaming experience. The use of eye tracking was also examined to see if it provides a means for continuous diagnostic monitoring of automation use and fatigue. Throughout the effort, the aim was to determine how different levels and reliabilities of automation moderate reliance on automation and the impact of fatigue. Specific objectives were as follows:

1. ***Test the Impact of Automation Configuration on Reliance, Trust, and Sustained Performance.*** Levels of automation (LOAs) refer to the tradeoff between operator control and

delegation of control to the machine. Higher LOAs (i.e., greater machine control) reduce operator workload, but also may result in vigilance decrements, loss of situation awareness, and complacency (Miller & Parasuraman, 2007). Another critical factor is the reliability of the automation. High reliability is desirable for performance optimization, but it may also encourage operator complacency. Building on existing work (i.e., Calhoun et al., 2011), this research investigated how fatigue impacts performance and over- or under-reliance on automation at different LOAs and reliabilities.

2. ***Establish the Operational Significance of Fatigue in Multi-RPA Control.*** Performance changes were examined during extended missions, using a multiple autonomous system simulation with several partially automated tasks presented during missions with varying levels of workload demands. The interplay between fatigue and automation is anticipated to vary with the context in which fatigue is experienced. Passive fatigue, linked to monotony, is distinct from active fatigue, linked to chronically high workload (Desmond & Hancock, 2001). Both fatigue contexts are potentially operationally significant. The aim of this objective was to distinguish their effects for envisioned multi-RPA operations.

3. ***Examine the Effect of Fatigue on Operator Reliance on Automation.*** Previous work (Neubauer, Matthews, Langheim, & Saxby, 2012) shows that fatigued surface vehicle drivers are more likely to use optional automation than those who are not fatigued. Thus, fatigue may encourage over-reliance on automation in the sense of voluntarily surrendering control to the machine. On the other hand, evaluating the utility of the automation may be perceived as a secondary task that the fatigued operator sheds, potentially leading to under-reliance. This research aimed to test whether fatigue experienced during multi-RPA operation contributes to over- or under-reliance, and whether any such effect depends on LOA and reliability of automation.

4. ***Explore the Role of Individual Differences in Reliance on Automation and Fatigue.*** Past research suggests that individual differences may influence operator interaction with automation. For example, video game experience has been demonstrated to impact performance in a simulated RPA task (Cummings et al., 2010). Given that skill typically protects the operator against fatigue-related impairment (Matthews, Davies, Westerman, & Stammers, 2000), this research tested whether video game experience protects operators against over-reliance while fatigued. Additionally, personality traits were examined in respect to automation interaction with the goal of identifying candidate personality drivers that influence reliance on automation. Additionally, this research extended the study of Szalma and Taylor (2011) that showed personality traits present differential responses to automation by determining the significance of fatigue in this respect.

5. ***Test the Utility of Eye Tracking Indices of Trust and Complacency.*** Few experiments have utilized gaze tracking to investigate trust in automation. However, several recent studies suggest how eye tracking could be employed to investigate the issue (e.g., Galesic, Tourangeau, Couper, & Conrad, 2008; Wickens, Dixon, Goh, & Hammer, 2005). These studies suggest that metrics such as the frequency of visual inspection of automated systems and average dwell time during operator interrogation of automation-provided action alternatives may provide objective indices of operator trust and reliance on automation. Eye tracking is

also useful as a fatigue index (Wierwille, Ellsworth, Wreggit, Fairbanks, & Kirn, 1994), but the utility of eye tracking for diagnostic monitoring in automated task environments has yet to be established.

6. ***Explore Manipulations of the Interface to Optimize Reliance on Automation.*** While industrial and educational psychology have examined the influence of individual differences on training and organizational structures, such measures have not been a main focus of human factors research. With advances in computing technology, it is now possible to dynamically adapt interfaces to take individual differences into account. For this objective the aim was to provide useful data on whether modifying the interface is likely to be successful.

4. BACKGROUND

Related to the specific research objectives delineated in Section 3, a more detailed review of pertinent literature pertaining to fatigue and individual differences is provided below.

4.1 Fatigue with Respect to Multi-RPA Control

Studies of RPA operators confirm that fatigue is a significant operational issue (Ouma, Chappelle, & Salinas, 2011; Tvaryanas & MacPherson, 2009). A variety of factors play a role, including those related directly to task demands, including long duty periods and human-machine interface difficulties (Ouma et al., 2011). However, there has been little work conducted on how fatigue responses may be controlled by task characteristics. The present effort aimed to establish the operational significance of fatigue in multi-RPA control.

Fatigue is a complex, multifaceted construct; different aspects of fatigue may vary in their impact on performance (Matthews, Hancock, & Desmond, 2012). As noted in surveys of RPA operators (Ouma et al., 2011), fatigue overlaps with stress, which is also multifaceted. Contemporary dimensional models of fatigue differentiate a variety of components including acute and chronic fatigue, and emotional, motivational and cognitive expressions of fatigue (Matthews, Desmond, & Hitchcock, 2012). The completed research followed the model proposed by Matthews et al. (2002), which differentiates fundamental state dimensions of task engagement (energy, concentration, task motivation), distress (negative mood, lack of perceived control) and worry (intrusive, distracting thoughts).

Use of a multidimensional model is important in research on fatigue and performance in order to characterize both the effects of the operational environment on fatigue, and to identify performance vulnerabilities. Recent work distinguishing active and passive fatigue states (Desmond & Hancock, 2001) demonstrates the need for a multidimensional perspective. Active fatigue develops in conditions of overload and frequent control operations, whereas passive fatigue is associated with underload and monotony. Studies of driver fatigue (Neubauer, Matthews, Langheim, & Saxby, 2012; Saxby, Matthews, Hitchcock, & Warm, 2007; Saxby et al., 2008) show that these two forms of fatigue correspond to different patterns of subjective state response. Passive fatigue is related to more rapid loss of task engagement, whereas active fatigue is characterized by distress and task disengagement. Furthermore, passive, but not active, fatigue relates to loss of alertness, operationalized as slowed response to an emergency event. In the RPA context, active fatigue may be most likely in periods of high task demands, requiring extensive multi-tasking, and passive fatigue may be typical of monotonous surveillance tasks.

Some relevant initial work on task-induced fatigue has been conducted at the University of Cincinnati. Guznov et al. (2011) showed that a simulated multiple robot control task tended to produce the high levels of workload and distress typical of active fatigue. A follow-up study (Guznov, Matthews & Warm, 2010) replicated this effect and showed that, in solo operators, subjective task engagement tended to decline over time. Guznov (2011) used a flight simulator to investigate the influence of workload on subjective state responses to performing a RPA mission requiring visual detection of ground targets. Task engagement declined substantially in a single-task condition, but to a lesser degree in a dual-task condition (vehicle operation + responding to audio messages). The state response seen in the lower workload, single-task condition resembled the passive fatigue response observed in vehicle driving studies (Saxby et al., 2007, 2008). Finally, another study compared the impact on stress and fatigue of several workload manipulations, including number of RPAs, time pressure, and feedback, using the "RESCHU" simulation (Donmez, Nehme, & Cummings, 2010).

Two types of cognitive impairment may be especially relevant in RPA operation, including loss of sustained attention and vigilance (Chappelle, McDonald, & King, 2010). A large literature on signal detection (Warm et al., 2008) demonstrates that vigilance is highly sensitive to fatigue. Work by Matthews and colleagues has shown that even mild-to-moderate levels of fatigue prior to performance of vigilance reliably predicts decreased perceptual sensitivity on a range of tasks requiring sustained attention (Matthews, Warm, Reinerman, Langheim, & Saxby, 2010a; Matthews et al., 2010b; Shaw et al., 2010). A recent doctoral dissertation (Guznov, 2011) confirmed that a measure of fatigue taken following training in target detection predicted subsequent detection performance in a simulation of RPA operation. The vigilance component of operations may be especially dangerous because vigilance tasks are often themselves a source of cognitive fatigue.

A further vulnerability to impairment originates in the multi-tasking nature of RPA operations. In addition to surveillance, operators must perform a variety of functions including vehicle control, communications and analysis of information from multiple displays (Mouloua, Gilson, & Hancock, 2003). Multi-tasking requires active, strategic management of the various task components. Such cognitive control is vulnerable to the demotivating effects of fatigue, as operators increasingly adopt effort-minimizing task strategies (Hockey, 1997), including task shedding (Schulte & Donath, 2011), and switching from a proactive to a reactive mode of control (Hockey et al., 1998; Sauer, Wastell, Hockey & Earle, 2003).

4.2 Relation of Fatigue on Operator Reliance on Automation

Potentially, automation may be beneficial in supporting the fatigued RPA operator. If some operator functions can be automated, operator impairments in cognitive control of multi-tasking may be mitigated. The operator's shrinking pool of processing resources may be focused on the subset of task components for which human involvement is critical, leaving more routine functions to the machine. Benefits may be especially pronounced in high-workload phases of operations that produce active fatigue (Desmond & Hancock, 2001). Similarly, automation may be especially helpful in next-generation multi-RPA operation.

However, there may also be dangers to automation technology, especially in circumstances that elicit passive fatigue. Generally, automation tends to shift operators from active controllers of their work activities to passive monitors of technology (Warm et al., 2008), implying a risk of task disengagement that may be exacerbated by fatigue. Specifically, the fatigued RPA operator may be inclined to mentally "coast" and let the machine do much of the work. For example, at the LOA

characterized by "management by exception," the machine chooses an action, which the operator can override within a limited time window (e.g., Liu, Wasson & Vincenzi, 2009; Ruff, Narayanan & Draper, 2002). Fatigued operators may become reluctant to perform the mental work necessary to decide that an override is necessary, i.e., becoming over-reliant on the technology.

The above-cited research suggests that in conditions of passive fatigue, the operator may show excessive trust or complacency, as an energy conservation strategy (Sauer et al., 2003). This is supported by observations in simulated vehicle driving research, in which fatigued drivers were more likely to engage vehicle automation, even though its use did not enhance performance (Neubauer et al., 2012). The impact of fatigue on trust may be especially pronounced when the automation is of high reliability. Conceivably, a different dynamic may operate when the automation is unreliable. The fatigued operator may be unwilling to exert the effort necessary to continually evaluate the utility of the automation, and so may ignore the automation where possible. Unreliable automation might also threaten the operator's sense of control over the system, elevating distress, and potentially exacerbating active fatigue. Ouma et al. (2011) implicated interface issues in burnout, consistent with this suggestion.

In sum, a case can be made that fatigue may augment some of the potentially harmful effects of automation. The stronger case is that fatigue, especially of the passive variety, may amplify over-reliance on automation, but in some instances fatigue might lead to neglect of automation. The research completed in this effort examined the effect of fatigue on operator reliance on automation.

4.3 Individual Differences in Video-gaming Experience

Usage of video games is becoming increasingly prevalent in adolescents and young adults (Anderson, Gentile, & Buckley, 2007). A recent review (Spence & Feng, 2010) concluded that video game exposure was positively associated with a range of sensory, perceptual, and attentional abilities that contribute to performance on tasks requiring spatial cognition. Benefits may be tied to the extensive practice that 'serious' gamers receive. Gentile (2009) reported that, in non-pathological gamers, the mean and standard deviation of weekly hours of game play were 11.8 and 12.6, respectively.

Could experience with video games build expertise that transfers to autonomous tasks such as RPA operation? Building a case of this kind requires results that show (1) commonalities in video gaming and RPA skills, (2) evidence that gaming enhances those skills, and (3) evidence that those skills transfer beyond the gaming task itself. Potentially, there are several types of skill that might generalize from gaming to RPA operation. Table 1 shows a simplified summary of critical aptitudes for RPA operations identified in Chappelle et al.'s (2010) analysis. Subject matter experts (SMEs) reported that operators lacking these attributes struggled to acquire relevant skills.

Table 1. Summary of Critical Aptitudes for RPA Operation (from Chappelle et al., 2010).

Attribute	Cognitive proficiency	Visual perception	Attention	Spatial processing	Memory	Reasoning
Example processes	Speed and accuracy of information-processing	Visual scanning, visual recognition	Vigilance, divided attention	Spatial analysis, spatial reasoning	Working memory, delayed memory	Problem solving, forward thinking, task management

Spence and Feng (2010) used a somewhat similar scheme to categorize the demands of different video genres, including action, driving, and maze/puzzle game types. Action games generally had the highest demands for attributes critical to RPA operations, including speeded processing, visual perception, and various forms of attention and spatial processing. They argued that 'first-person shooter' games (e.g., *Halo, Call of Duty, Battlefield*) were probably the most prevalent type from the action genre.

A variety of cognitive functions correlate with level of exposure to video games (especially action games), including visual search, visual attention, visual memory, contrast sensitivity, and mental rotation (Richardson et al., 2011; Spence & Feng, 2010). As these are correlational findings, they might reflect the influence of a third variable; perhaps individuals attracted to video gaming have high initial spatial ability, for example. Experimental studies have examined directly the effect of training on cognitive functioning. The majority of such studies confirm that training on action games improves aspects of spatial or attentional functioning (Green & Bavelier, 2008; Richardson et al., 2011; Spence & Feng, 2010).

Importantly, some studies provide evidence for 'far transfer;' training on video games produced improvements on other spatial tasks that are not closely similar to the task used for training (Spence & Feng, 2010). For example, Spence, Yu, Feng and Marshman (2009) recruited participants with no previous first person shooter gaming experience. Training on the *Medal of Honor: Pacific Assault* game improved performance on a spatial attention task using stimuli unrelated to the game. Practice on action games (Spence & Feng, 2010), as well as on tasks requiring working memory or executive control (Tang & Posner, 2009), may enhance functioning of brain networks for attention.

Thus, practice on action games may indeed enhance skills that are relevant to RPA operations. In an RPA simulation study, McKinley, McIntire, and Funke (2011) confirmed that experienced gamers showed visuospatial attention skills that exceeded those of pilots, and matched pilots in aircraft control skills. What remains to be investigated is the extent to which games improve the specific skills needed for optimal usage of automation. One relevant study is Cummings et al. (2010) who found that experienced gamers collaborated more effectively with automation in a simulated RPA task. However, it remains to be shown that gaming experience protects against adverse effects of fatigue. Broadly, skilled operators are less vulnerable to fatigue and stress effects than those of lesser expertise (Matthews et al., 2000), but the influence of skill on sustaining effectiveness in the RPA context remains to be explored. Should the role of video gaming expertise be substantiated from this effort's research, there is a justification for further work on using game-like environments for training and selection of experienced game players as a potential mitigation strategy.

4.4 Individual Differences in Personality

Recent research has illustrated the importance of attention to individual differences, such as personality, when considering automation reliance and the role of trust (Burkolter, Kluge, Matthias, 2010; Szalma & Teo, 2010). For instance, Merritt and Ilgen (2008) found that when automation characteristics were held constant in an X-ray screening task, participants' perceptions accounted for 52% of the variance in trust in automation. Moreover, extraversion was the individual difference measure noted to be related to propensity to trust and it moderated automation reliance.

In regards to research paradigms targeting supervisory control of autonomous vehicles, individual differences are now more frequently being examined. Chen and Barnes (2012) examined differences in UGV control in terms of spatial ability and perceived attentional control (PAC). Participants with high spatial ability performed tasks that involved switching between remote and local terrain views more effectively than those of lower ability (Chen & Terrence, 2009). In a different ground vehicle simulation, Szalma and Taylor (2011) examined the relationship of operator personality (Five Factor Model) and operator performance, workload, stress, and coping. The results showed that all five traits were associated with differences on at least one measure of perceived workload and stress. However, the pattern of relationships between traits and dependent variables varied and task characteristics exerted the strongest influence. Focusing on RPAs, Cummings and colleagues (2010) categorized their simulation participants into three groups: automation consenters, dissenters, and mixed consenters. The dissenters often chose to ignore automated re-plan reminders. This tendency hurt their performance and prompted the authors to postulate that personality type can influence participants' performance.

Other RPA simulation-based research has shown differences associated with personality. Using a post-test procedure similar to Cummings et al. (2010), participants categorized as high achievers tended to distrust the automation, resulting in inflated task completion time (Ruff & Calhoun, 2011). Replicating the experimental paradigm, personality measures were recorded in a follow-on experiment and results (Calhoun, Ruff, & Murray, 2012) suggested a complex interplay between personality factors, level of automation, and task type. For instance, participants high in neuroticism tended to perform better on an infrequent change detection task. Performance on a frequently occurring image analysis task was better for participants showing low extraversion. In addition, low levels of autonomy were associated with better performance for participants who were high in neuroticism, conscientiousness, and agreeableness. A third experiment included a condition in which participants could change which of three intermediate autonomy levels was in effect (Kidwell et al., 2012). With this adaptable control scheme, participants' choice of autonomy level was a behavioral indicator of reliance on automation. Results showed a very strong correlation between autonomy level choice and extraversion: highly extraverted participants chose the highest level of autonomy, which only required them to respond if they wanted to veto the automation's recommendation. In contrast, less extraverted participants chose a level of automation that required the operator's consent before acting. These results demonstrate the utility of a multi-RPA simulation for exploring the role of individual differences in interacting with automation.

5. RESEARCH APPARATUS

The experimental apparatus employed to gain a better understanding of the circumstances under which individual differences, fatigue, and automation characteristics may interact to produce inappropriate reliance on automation consisted of a multi-RPA simulation and an eye tracker. Both components underwent considerable modifications throughout the course of this effort to meet experimental objectives. The following provides further details.

5.1 Multi-RPA Simulation

The multi-RPA simulation was a system referred to as "ALOA" (Adaptive Levels of Autonomy; version 3), an automation research test bed developed by OR Concepts Applied (ORCA; Johnson, Leen, & Goldberg, 2007). This simulation incorporates the ORCA commercially

available routing software/mission planner to provide needed complexity and realism. The tasks supported by the simulation are designed to represent the cognitive task demands envisioned for a single operator supervising multiple autonomous aircraft in a full mission scenario. Participants' interaction with the simulation consists of monitoring the displays and making inputs via mouse and keyboard.

The simulation software was executed on a custom iBuyPower Workstation with an Intel® Core™ i7-4820K CPU processor @ 3.70 GHz, 16.0 GB RAM, and a nVidia GeForce GTX 770 graphics card (Microsoft© Windows 7 Enterprise 64 bit Operating System). Two monitors provide numerous windows to support participants' completion of multiple tasks while supervising the automated flight of three aerial vehicles. A keyboard and mouse are used for participants' inputs.

5.1.1 Baseline ALOA Multi-RPA Simulation

Figure 1 shows a screen shot of the baseline ALOA simulation that has been used in numerous evaluations (e.g., Calhoun et al., 2011). Note the annotations highlighting each task window. A unique feature of the simulation is that for three of the primary tasks highlighted in Figure 1, the LOA can be manipulated independently and range from manual to fully autonomous levels. In the current effort, two intermediate LOAs were used as is customary for automated decision aids: manage by consent and manage by exception. Additionally, the experimenter can differentially establish the automation reliability of each task. The simulation also has numerous secondary tasks that can be included in experimental trials to induce different workload levels. The software supports data recoding for each task (task completion time and accuracy in most cases). The ALOA simulation is also unique in that there are three different control schemes available for setting LOA within the task: static (established by the experimenter and fixed for the entire trial), adaptive (based on participant performance, mission event, or experimenter manipulation), or adaptable (under participant control). Unless indicated otherwise, the LOAs for tasks were static for this effort's experiments.

The ALOA simulation also supports multiple experimental configurations. The script editor interface allows the experimenter to define the number of mission events and the nature of their occurrence. Events can be programmed either by specifying their frequency of occurrence or by stipulating specific times of occurrence (after mission start) for each desired event. Missions can vary in the number of task types presented, the frequency of each task type, and the timing of each task. For the image analysis task, the experimenter controls what is displayed in the task's window, ranging from a simple photo with overlaid geometric symbols (e.g., to support an easily trained "count the number of diamonds" task) to a moving video (requiring the detection of a target of interest). In fact, the workload and difficulty of several of the simulation's tasks can be manipulated. More information is provided below on how each task type (see Figure 1's annotations) was implemented in the present effort.

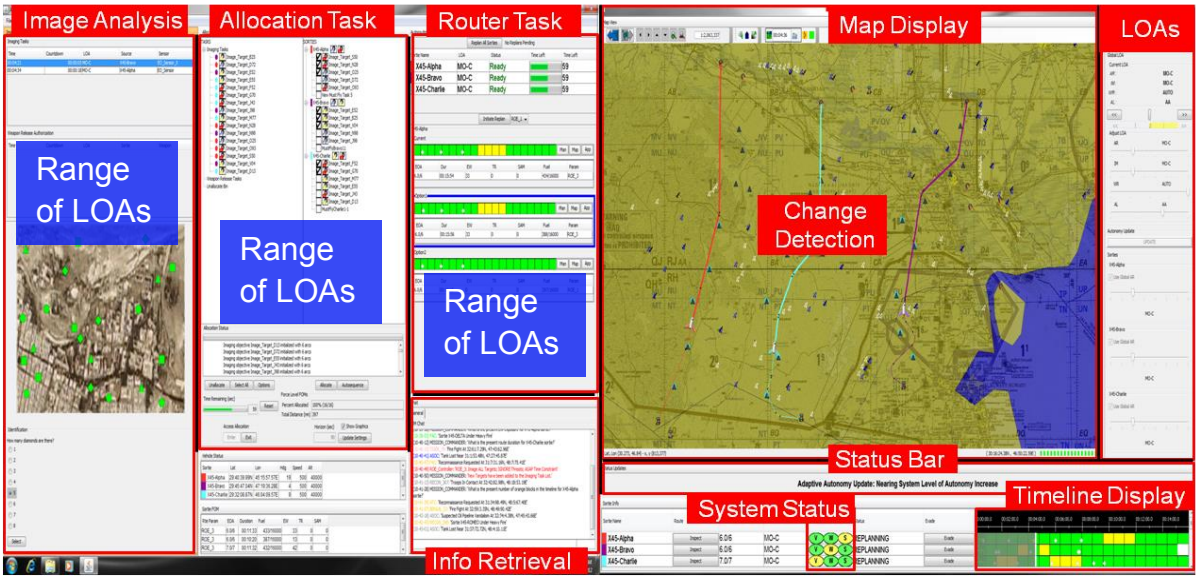


Figure 1. Baseline Adaptive Levels Of Autonomy simulation (ALOA) annotated screenshot.

Allocation and Router Tasks. Participants were instructed that these two tasks were the highest priority and should be completed, in tandem, as quickly as possible, whenever they started (signaled by an aural "ding" and a text message in the status bar). It was explained that failure to do so, would result in missing targets and routes failing. These instructions, in addition to employing a very high LOA for these tasks, helped ensure that each trial was properly executed (i.e., the routes passed the necessary targets which prompted surveillance tasks that were key to examining participants' reliance on automation). With the high LOA, participants were only required to click on a button to start the automated process that allocated RPAs to any new unassigned targets (indicated by the blank circles). The automation then accurately assigned vehicles to targets such that each was viewed by an RPA with the appropriate sensor type (indicated by color). Participants were instructed to monitor the allocation progress (the termination of which was indicated by a "100% complete" readout).

Once the allocation was completed ("100%" displayed), the participant waited for the automation to present a "Ready" message for each aircraft listed at the top of a rerouting window. When the ready message appeared for each aircraft, the participant was to click on each aircraft in the list and choose one of two new routes presented in the section below the vehicle list (the current route was listed above the two new options). Each new route had different ROE (rules of engagement) indicated (i.e., the ROE for which the route was optimized) and the participant was instructed to choose the route whose ROE matched the mission commander's ROE instruction from the chat window. The ROE was assigned based on the current needs of the mission and each ROE (numbered 1-3) emphasized different criteria (e.g., "ROE_Controller: ROE_1: Image ALL Targets; AVOID Threats; NO Time Constraint").

Image Analysis Task. In the present effort, this surveillance task was the second priority task type, after allocation and router tasks. Figure 2 shows the task window of the image analysis task in the baseline ALOA simulation. Participants were prompted that an image taken by an aircraft was waiting to be analyzed by the addition of a row in the image task window (top queue) that included an identifier, time added, vehicle source, and counter showing analysis time remaining. Symbology in a timeline below the map (Figure 1) also provided cues of pending

images (though timeline functionality was removed after the first study to avoid time tracking by participants). Participants had 30 s to complete the analysis before the image blanked and the task was recorded as a 'miss.' Task completion began with row selection that called up a photo with 19-26 overlaid green shapes (diamonds, squares, circles, and triangles). Participants were to count the green diamonds and choose an option below the image (1-8) that matched their count. An automated decision aid (with medium LOA) recommended one of the options by highlighting the option and filling in its radio button. To complete the task and clear the photo, participants needed to click a "Select" button at the bottom of the window if the LOA was manage by consent. If the LOA was manage by exception, participants could click "Select" after choosing an option, or simply allow the countdown timer to expire to make a selection. Thus a participant could simply allow the automation to perform the task without interference.

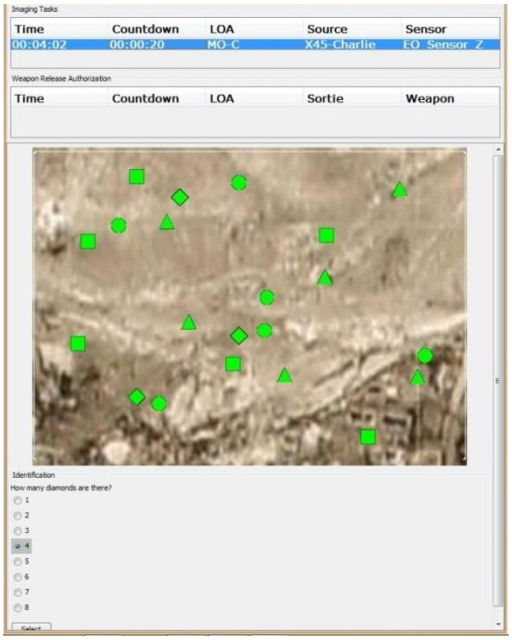


Figure 2. Sample close-up of image analysis task window.

Change Detection Task. Periodically, a red plane symbol appeared at a random location on the map. Participants were instructed to click on the symbol with the mouse (alerting "headquarters" of the presence of the unidentified potential threat). This was the next priority task, after the Allocation, Router, and Image Analysis task types. Completion of this step (or the expiration of 10 s) caused the symbol to disappear. Completion time was measured from the appearance of the symbol until the participant clicked on the symbol. The number of misses was also recorded.

Two other tasks available in the baseline ALOA system were deemed to be lowest in priority. These tasks help represent anticipated demands on a supervisor of multiple RPAs, as well as impose workload. Participants' shedding of these lower priority tasks could be viewed as an indicator of fatigue state.

Information Retrieval Task. Participants were required to monitor the chat window for requests for information to be retrieved from various testbed windows (RPA's heading, fuel level, current number of objectives, route duration, threat exposure, SAM exposure, imaging tasks

remaining, or number of yellow, red, or black blocks in the timeline). For each information request, participants typed the corresponding value and selected 'Enter' in the chat window. Speed and accuracy in responding to the information requests were measured but not analyzed. For experiments conducted in this effort, the ratio of chat messages related to tasks (information retrieval and ROE changes) to those not task related ("chatter") was 15:40. Also, only the ROE messages appeared in red, increasing the participants' workload to discern chatter messages from information request messages.

Systems Status Task. In a window below the map (Figure 1), the status of the vehicle ("V"), weapons ("W"), and sensors ("S") for each RPA was indicated. Periodically, the color of one of the circles changed from green to yellow (indicating a change from normal to abnormal status). Participants were required to monitor this information and click on the circle with the mouse to indicate awareness of the problem. If the button wasn't clicked within 10 s, it changed from yellow to red and remaining in the red state for another 5 s unless clicked. Thus, participants had 15 s to fix each failure, before the task was counted as a miss. The circle color returned to green with participant selection or after the 15 s timeout period. Response time was measured from when the circle changed to yellow until the participant clicked the circle.

5.1.2 Enhanced ALOA Multi-RPA Simulation

A significant portion of the first year of this effort involved re-configuring the ALOA multi-RPA simulation to better meet the research objectives listed in Section 3. The baseline simulation was designed to run repeated trials approximately 10-20 minutes in duration with a task load of approximately 4 tasks/minute. Repeating the available short trials/scenarios to fill more lengthy sessions was deemed inadequate in inducing fatigue as 'participant states' would reset during the breaks required to re-initialize the simulation. Rather, trial duration needed to be increased to support the investigation of the effects of sustained operations on participants' use of automation. Despite several technical difficulties given ALOA's use of a commercial router, two trials of 60-minute durations were generated, each with a different but similar scenario (e.g., defined locations and routes for RPAs and targets that were imaged).

Besides trial length, the types of tasks provided in the baseline configuration were revisited. All the task types described in Section 5.1.1 were retained. However, several types of tasks were added to provide data to help inform the research objectives. These tasks are described in detail below and are reflected later in Figure 5).

Digit Pairs Task. A task type with a sufficiently high 'event rate' to facilitate the detection of any loss of vigilance during the period of work was added, given that vigilance is likely to be the most fatigue-sensitive performance indicator. (Broadly, it was anticipated that passive fatigue would produce greater loss of vigilance than active fatigue (cf., Saxby et al., 2008), whereas active fatigue would produce greater strategic efforts at reducing workload, e.g., by task shedding.) The cognitive vigilance task added to help measure a temporal change in performance was based on a "digit pair task" employed by Bolia, Nelson, Middendorf, Guilliams, and McLaughlin (2004). It involves deciding if presented digit pairs meet requirements of a critical signal. To implement this task in ALOA, three small boxes (digits, 'True', & 'False') periodically appeared, overlaying the map (see Figure 3). The participants' task was to compare the two digits and determine if the two numbers were within one digit of each other (or the same number) and if the sum of the two numbers was between '3' and '15' (and not '3' or '15'). If both criteria were met, participants'

correct response was to click "True" with the mouse. If at least one criterion was not met, participants' correct response was to click 'False.' The three boxes disappeared after a response was clicked, or the task timed out (e.g., 10 s). The probability of any given stimulus presentation being a critical signal was approximately 80%.

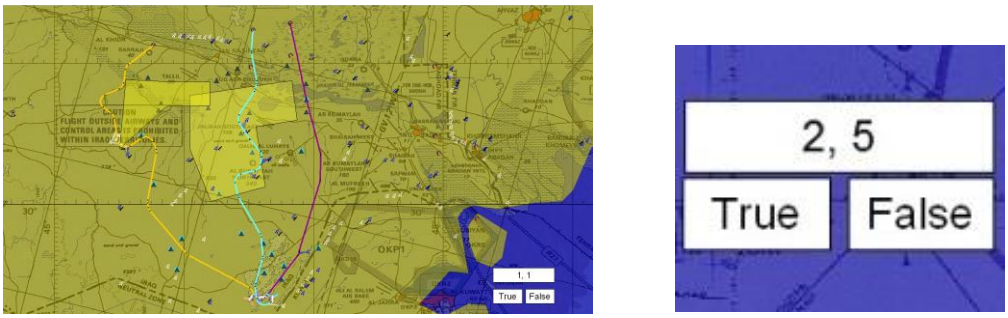


Figure 3. Sample close-up of digit pairs task that overlaid the map.

Communications Task: A second task type added to the baseline ALOA simulation required participants to monitor an audio stream to represent operational aural workload. This task type also avoids any modality interference issues associated with the other visual-based ALOA tasks. A version of the Coordinate Response Measure (CRM), a communication performance task, was employed (Bolia, Nelson, Ericson, & Simpson, 2000). The CRM elicits multiple radio call recordings from 4 speakers, each including a specific call sign followed by a combination of one of four possible colors and a number from 1-8 (e.g., "ready *Eagle*, go to *blue 8*"). The auditory load was such that participants heard eight different call signs (including "Eagle") during the trial according to the frequency setting for the particular study. When participants heard "Eagle", they entered the first letter of the color and the number into the chat window. For example, for the call "ready Eagle go to red-six now", the participant typed "r6" and pressed "Enter." Data from pilot trials indicated that the CRM task increased workload, without significantly increasing time to train participants.

Weapons Release Task. To augment the image analysis task available in the baseline configuration, another surveillance task was added to facilitate measuring operator reliance on automation. To revisit, over-reliance may take the form of neglecting to check the accuracy of automated functions. Trust in automation increases with reliability, but increasing trust may also encourage over-reliance on automation or complacency (Parasuraman & Wickens, 2008). Several studies (e.g., Dixon, Wickens, & McCarley, 2007; Rice & Keller, 2009) have indexed over-reliance from the operator failing to override errors made by automated systems. Dixon et al. (2007) introduced metrics that distinguish between reliance (operator failure to correct automation 'misses') and compliance (operator failure to correct automation 'false positives'). Excessive reliance and compliance represent different forms of deviation from appropriate levels of trust. In the present research, reliance was indexed more generally as the extent of operator agreement with the automation.

The added surveillance task involving authorization of a weapons release was designed to generate metrics within the ALOA simulation similar to those employed by Dixon et al. (2007). As illustrated in Figure 4, measures of hits, misses, false alarms, and correct rejections can be measured with this task. Participants are signaled that there's an image to be analyze by the addition of a row at the top of the weapon release authorization field of the imaging tasks window.

Clicking the row called up an image of terrain with several tanks (instead of the image with overlaid geometric shapes used for the image analysis task; see Figure 5). Participants were tasked to analyze the tanks and determine which were hostile versus friendly, according to their appearance. Friendly tanks had shorter barrels and a wider body in contrast to enemy tanks that had longer barrels and a thinner body.

Some of the tanks were highlighted with a red box showing which tanks an automated target recognizer had identified as hostile. (The accuracy in which the automation marked images with red boxes was driven by the reliability setting for the particular study.) Participants could not modify the assignment of red boxes. They either confirmed or rejected the conclusion of the automated targeting system by authorizing a strike when all hostile tanks were highlighted, or not authorizing a strike otherwise. Thus, this task was designed to measure how much participants relied on the automation (red boxes), as the images were purposely made difficult to analyze (by adjusting blur and color) to better examine responses reflecting reliance and participants' stress levels. (Participants were briefed that "sometimes the image will become temporarily obscured with static to represent the unreliability of the data link connection.")

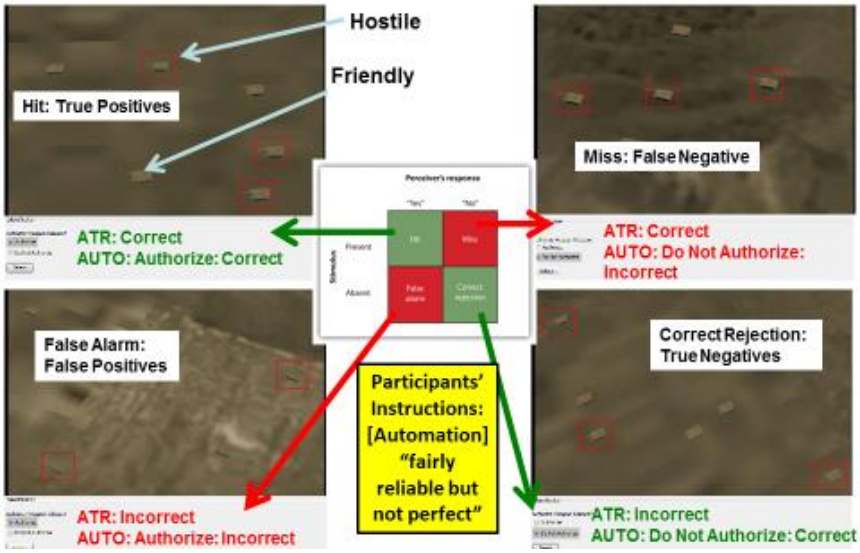


Figure 4. Sample close-up of weapons release authorization task.

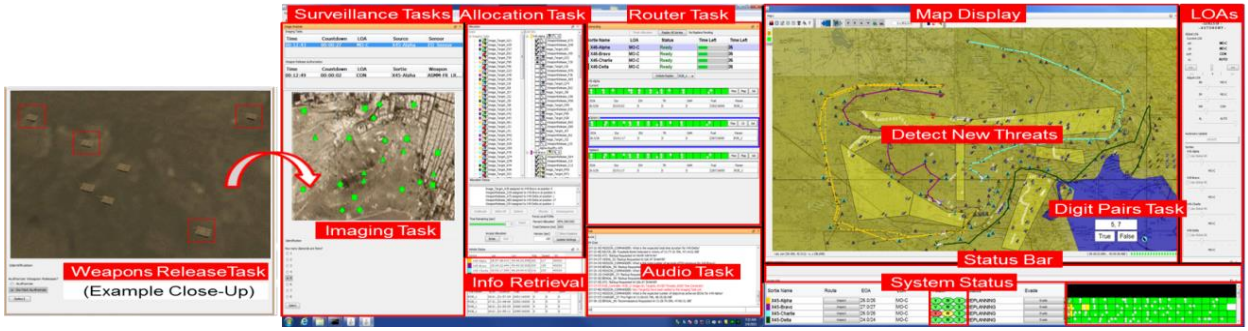


Figure 5. Enhanced Adaptive Levels Of Autonomy simulation (ALOA) annotated screenshot.

5.2 Eye Tracker

For some experiments, eye movement data were collected using a faceLAB eye tracker (Seeing Machines Inc., version 5.0, Figure 6). The desktop-mounted system consists of two infrared cameras and a group of infrared light emitting diodes. The system measured infrared corneal reflectance and pupil and head position to determine point of gaze at 60 Hz. While the system automatically outputs a wide variety of eye gaze metrics, the most relevant to this research include frequency and duration of eye blinks, percentage of eye closure, and frequency and duration of fixations and saccades. Eye tracking was recorded on a separate system from the simulator on a Hewlett-Packard z400 Workstation with an Intel® Xeon® W3565 processor @ 3.20 GHz, 12.0 GB RAM, and an AMD FirePro V4900 (x2) graphics card (Microsoft© Windows 7 Professional 64 bit Operating System).

Figure 6. faceLAB Eye Tracking System.

5.3 Eye Tracker Link with ALOA Multi-RPA Simulation

One objective of this effort was to explore adaptations of the interface to mitigate the impact of fatigue and optimize reliance on automation (Section 3). Previous experiments exploring physiological indices for adaptive automation have frequently relied on metrics derived from electroencephalography (EEG) and event-related potential (ERP) analysis (e.g., Mikulka, Scerbo, & Freeman, 2002; Pope, Bogart, & Bartolome, 1995; Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2003). However, eye tracking methods may provide substantial benefits over other methods for assessment of participant interaction with an automated system, in that eye tracking is less invasive (table mounted systems require relatively minimal calibration, without electrodes attached to the participant) and may allow diagnosis of changes in participant interactions with automation (i.e., shifts in compliance, reliance, and monitoring). This would involve data from the eye tracker being fed to the ALOA simulation to trigger one or more interfaces to adapt in some manner (e.g., change the LOA of a certain task or present a message in the chat window).

To integrate the eye tracker with the ALOA simulation, it was determined that the easiest and most flexible approach would be one that allows each system to perform completely separate processes that communicate with one another. In other words, the communication would be applied to trigger changes in one or more ALOA interfaces based on on-going analysis of the participant's gaze point and other eye tracker data in the computer associated with the eye tracker. The ALOA developer, OR Concepts Applied, designed and implemented a messaging protocol in

ALOA for external processes to change the LOA, obtain current status, and display system status updates and chat messages.

Further detail about the integration is presented below in Section 8.3. Study 3. Diagnostic Fatigue Monitoring for Adapting Level of Automation.

6. RESEARCH OVERVIEW

This research involved extensive collaboration of AFRL and scientists in academia. The linked efforts also provided an avenue to leverage the organizations' respective experimental resources, namely, AFRL's multi-RPA simulation and eye tracker/ocular-related algorithms and academia's large subject pools that facilitate examining individual differences data. Both AFRL and academia used identical test apparatus in support of this research effort.

6.1 Simulation Configuration and Pilot Testing

Initially, the focus of the effort was on developing ALOA task configurations suitable for investigating reliance optimization under fatigue (see Section 5). This was followed by pilot testing conducted at the University of Cincinnati (n=6) and the University of Central Florida (n=12) to confirm that the configured tasks and training paradigm were at an appropriate difficulty level for the anticipated student participant pool. The pilot research also served to finalize how the new experimental tasks should be configured as well as the frequency of each task type for the test scenarios employed. Levels and reliabilities of automation to employ in the research were examined in addition to the objective and subjective measures. A key concern in the pilot testing was to validate task configurations for inducing the build-up of active and passive fatigue over time (similar to Saxby et al., 2007). One element of validation was to show that the manipulations produced the contrasting patterns of state change characteristic of active and passive fatigue (Saxby et al., 2007, 2008), with the goal of identifying the minimum duration adequate for the research objectives to make the conduct of the experimental test sessions across a large participant subject pool more manageable. This involved multiple cycles of experiment refinements after instances of pilot data collections. Refinements included changes to the simulation software, scenarios, training, experimental procedures, and data logging routines.

6.2 Study 1. Workload and Level of Automation Effects

Once pilot testing was completed, experimentation commenced at the University of Central Florida to examine the role of individual difference factors associated with personality, trust and video gaming experience in automation usage during sustained operations. Specifically, this first full-scale experiment focused on how active and passive fatigue effects on performance and automation reliance vary with LOA. Passive fatigue is more damaging to attention because the operator becomes disengaged from the task (Desmond & Hancock, 2001; Matthews, Hancock et al., 2012). It was expected that passive fatigue effects would be accentuated by use of a higher LOA, encouraging the fatigued operator to rely increasingly on the automation. Conversely, active fatigue effects are mediated by excessive workload, so higher LOAs were thought to be beneficial under those circumstances because they should reduce workload. Task load was manipulated in Study 1 to induce contrasting states of passive fatigue ("monotony") and active fatigue ("prolonged overload"; similar to previous work on simulated vehicle driving; Matthews, Neubauer, Saxby, & Langheim, 2012; Neubauer et al., 2012). This involved constructing two 60-

minute trials that differed on the number of each task type. Generating different trial scenario configurations was based on the assumption, for example, that passive fatigue may be induced using underload conditions, such as controlling a single RPA with few in-flight events. Conversely, active fatigue should ensue when the operator controls multiple RPAs and must attend to numerous events under time pressure. This experiment also examined LOA and involved measurement of task accuracy, reliance, stress, fatigue, and workload. Participants were 43 men and 58 women (mean age = 18.95 years). Section 7.1 describes key findings from this study; more detail is provided in Lin et al. (2015).

6.3 Study 2. Fatigue and Reliability Effects

This study tested how the effects of LOA and fatigue identified in Study 1 vary with reliability, in a lower-workload configuration. As automation reliability decreases, the operator is likely to allocate increasing effort to monitoring the automation for errors. This process is potentially damaging in active fatigue conditions (risk of overload of attention if effort has to be diverted to monitoring the automation), but helpful in passive fatigue conditions (helping the operator remain proactive rather than reactive). To better ensure that participants reached the desired fatigue state and to confirm if fatigue influences suboptimal use of automation, it was decided to increase the trial scenario duration to 120-minutes, as well as decrease the event frequency of some task types. Thus, this study was designed to determine the impact of automation reliability on performance, reliance, and fatigue using a scenario designed to induce passive fatigue with low task load. Participants were 81 men and 50 women (mean age = 19.86 years). The same measures utilized in Study 1 were employed. Study 2 findings are summarized in Section 7.2 and further detail is given in two additional articles (Lin et al., 2016; Wohleber, Calhoun et al., 2016). This study also employed an eye-tracker to begin examining the utility of ocular parameters for measuring fatigue and for diagnosing inappropriate operator reliance, compliance, and monitoring of the ALOA automation. Data were collected from a subsample of Study 2 participants: 26 men and 13 women (mean age = 19.86 years).

6.4 Study 3. Diagnostic Fatigue Monitoring for Adapting Level of Automation

The objective of this study is to explore how real-time eye parameter recordings might be employed to drive interface adaptations and improve task performance. In fact, research utilizing eye gaze behavior for adaptive automation is represented infrequently in the scientific literature and what has been conducted is almost exclusively focused on detection and mitigation of operator workload. However, this research is relevant in its demonstration that even relatively low level indices of eye gaze behavior such as pupil diameter (de Greef, Lafeber, van Oostendorp, & Lindenberg, 2009) and fixation dispersion (Fidopiastis et al., 2009) can successfully actuate adaptive automation, lending credence to the suggestion that it could be applied in the current context as well. In Study 3, eye fixation-based thresholds, determined from data collected in Study 2, are being applied for detection of changes in fatigue state that then trigger changes in surveillance task LOA. More details are provided in Section 7.3.

7. OVERVIEW OF RESEARCH PROCEDURES

Section 8 provides more details on each study, including citations for further information. Here, a general overview of the procedures is provided.

7.1 Participants

College students participated in the studies for course credit. Participants represented the age group and educational level of the military service core that may be selected for future RPA operations. All participants confirmed that they were fluent in English and had normal hearing, and normal, or corrected to normal, 20/20 vision. Medical conditions, including psychiatric conditions, which may compromise participant safety or performance, were grounds for exclusion. Individuals taking psychoactive drugs were excluded. None were experienced pilots.

7.2 Procedures

Sessions began with participants' informed consent and a brief overview. Next, participants completed questionnaires to measure demographics, individual differences, and subjective state (see Section 7.4). Training followed with an explanation of the ALOA simulation's displays and controls. The automation was described as "somewhat reliable, but prone to error" when automation was 60% reliable, and "quite reliable, but not perfect" when the automation was 80% or 86.7% reliable. Next, each task type was described and practiced, in turn, in single task vignettes. This was followed by a 15-minute training trial where participants were required to complete all the task types. The training trial was repeated if participants failed to complete the primary tasks accurately and within the system defined time-out limits. Training took approximately 30 min and was followed by a single experimental trial (either one 60-minute or 120-minute trial) with the assigned experimental condition (e.g., workload level, automation level/reliability, and static/adaptive LOA). All experimental trials contained the same types of tasks (see Section 5). Also, all trials used the same task prioritization scheme: top priority - allocation and router tasks; second priority - image analysis and weapons release task; third priority - change detection task. The remainder task types were lower and equal priority: information request task, systems status task, 'digit pairs' task, communication task. However, the frequency of each task type differed across experiments depending on the trial length and whether the objective was to induce passive or active fatigue. Tasks of particular interest to examining participants' reliance on automation (the image analysis and weapons release surveillance tasks) were presented at the same frequency across conditions within each study. The entire session time was approximately 2-3 hours per participant.

All data were analyzed qualitatively and quantitatively using descriptive statistics. Objective performance data were analyzed with parametric statistical techniques (repeated measures Analysis of Variance, ANOVA). Mixed model ANOVAs were typically used to test for statistical differences in subjective states. Correlation analyses were also applied to examine the relation of individual difference and experimental variables.

Given the focus on individual differences, power analyses for the correlational analyses were conducted for Study 1 (n=101) and Study 2 (n=131). Conventionally, Pearson r values of 0.1, 0.3, and 0.5 define small, medium and large bivariate associations. A study design should have adequate power to detect medium or larger magnitude Pearson rs, given that small rs are rarely of practical significance. In the power analyses, Type I error probability α was set to 0.05 (two-tailed). For Study 1, power values for small, medium and large r values were calculated as 0.17, 0.88, and 1.00. For Study 2, these values were 0.21, 0.94, and 1.00. Thus, power was adequate (>.80) to detect medium magnitude correlations in both studies. Studies were not intended to test for small magnitude correlations.

7.3 Objective Data

To address the key objectives of the studies, data analyses typically focused on three performance metrics for each surveillance task (image analysis and weapon release authorization): accuracy, reliance, and neglect. Accuracy was defined as the percentage of correct responses. Reliance on automation was measured by the percentage of total responses on which the participant followed the recommendation from the automation. The frequency of surveillance tasks not initiated (i.e., the row was not clicked by the participant to call up the image for the task) was considered a measure of neglect.

7.4 Subjective Data

Prior to training and objective data collection, the following instruments were administered:

Video Gaming Survey. As part of a demographic questionnaire, participants were asked to assess their video gaming experience and expertise overall, for both first person shooter games and for action games (e.g. "Estimate your level of expertise playing video games, in general"). Participants responded to experience questions based on an 8-point scale indicating number of hours a week anchored by 0 = "0-1" and 7 = "20+". Participants responded to expertise based on a 7-point scale (0 = "no expertise" and 6 = "expert"). Items were selected on the basis of existing research that suggests the importance of distinguishing different forms of gaming (Spence & Feng, 2010).

40 Mini-Marker Personality Scale. This scale, derived from Goldberg's (1992) 100-item scale personality scale, assessed Big Five factors (Saucier, 2002) of extraversion, agreeableness, openness, conscientiousness, and neuroticism. Participants used a 9-point Likert scale anchored by 0 = "Extremely Inaccurate" and 9 = "Extremely Accurate", to rate how well 40 different adjectives described their personalities.

One instrument was administered both before and after data collection:

Dundee Stress State Questionnaire. States such as active and passive fatigue (Desmond & Hancock, 2001) differ qualitatively; therefore, a multidimensional scale that gauges three higher order dimensions of subjective state related to self-regulation during performance was employed. The Dundee Stress State Questionnaire (DSSQ: Matthews et al., 2002) was used to assess intra-task changes in relevant state dimensions such as task engagement, distress and worry. Task engagement was of particular interest, as it has been found to be sensitive to passive fatigue manipulations (e.g., Saxby et al., 2008). Participants responded to statements describing current emotional state using a 5-point Likert scale anchored by 0 = "Definitely false" and 4 = "Definitely true." Both pre-and post-task (latter: instructed to consider trial's final 10 minutes) versions were recorded to assess stress state in response to task performance.

After objective data collection, subjective data were collected with the following measures:

NASA-TLX Task Load Index. This workload assessment tool (Hart & Staveland, 1988) provided a multi-dimensional rating procedure that addressed the mental, physical, and temporal

demands of a task, in addition to the participant's performance, effort, and frustration. Each of the six dimensions was assessed with a twenty-step bipolar scale. Analysis employed the raw scores.

Human-Computer Trust Scale. The human-computer trust scale (HCT; Madsen & Gregor, 2000) is a reliable (Cronbach's $\alpha = 0.94$) measure of affective and cognitive components of trust in automation. Items gauge confidence in an automation and willingness to act on the automation's recommendations through five constructs which are believed to impact level of trust in an automated decision aid: perceived reliability (R), perceived technical competence (T), perceived understandability (U), faith (F), and personal attachment (P). This study used a shortened version with items from each of the five constructs (R3, F3, T3, U2, R4, F1, P4, & U3). Participants responded to items based on their experience with the decision making aid automation by indicating the extent to which they agreed with statements about their trust by using a 5-point Likert scale anchored by 0 = "Extremely disagree" and 4 = "Extremely agree" (e.g. I can rely on the system to function properly).

Experimenter-generated Form. Participants completed a form assessing trust and automation usage with questionnaire items that pertained to specific ALOA components, based on previous work at AFRL.

7.5 Eye Tracker Data

For some studies, ocular parameters were recorded to determine their utility for detecting fatigue state as well as inappropriate reliance on automation. Specifically, the following were recorded: fixation frequency and duration, blink rate and duration, percent time eye closed (PERCLOS), and gaze point. Additionally, fixation type was determined based on Schleicher, Galley, Briest, and Galley's (2008) parameters for express (<150 ms), cognitive (150-900 ms), and overlong (>900). Frequency of each fixation type was tallied and the percentage of the occurrence of cognitive fixations relative to all fixations was calculated. For more information on specific metrics see Wohleber (2016). Criteria for inclusion were necessarily loose to accommodate limitations with the eye tracking hardware's ability to track edges. For example, researchers attempted to meet a standard of mean angular error of 2 degrees or less, however, data was still recorded for participants if this angular error requirement was not met along the left and right edge of the dual monitors. For more on inclusion criteria see Wohleber (2016).

8. SPECIFIC STUDIES

For each experimental study conducted, the following provides a brief overview of key methodological details and results. Further information is available in the cited publications.

8.1 Study 1. Workload and Level of Automation Effects

The first full-scale experiment focused on examining active and passive fatigue effects on performance and automation reliance as a function of the LOA. Additional detail for Study 1 is available (Lin et al., 2015). Two different 60-minute trial scenarios were implemented to induce, respectively, passive and active fatigue. The scenario designed to induce passive fatigue employed a mean task frequency of 4.8 tasks/minute. This is similar to what was used in previous research (15-minute scenarios, 6 tasks/min; Calhoun et al., 2011). In contrast, the task frequency was

increased significantly in the scenario designed to induce active fatigue (mean of 16.7 tasks/minute). Table 2 provides details pertaining to each of ten task types. The five task types at the bottom of the table differed in frequency between the low workload (passive fatigue) and high workload (active fatigue) trials. In contrast, the frequency of other tasks was held constant across scenario types, to better compare performance in terms of reliance on automation.

Table 2. Task Information including Task Frequency in Passive and Active (Low and High Workload) Scenarios.

TASK	60 MIN-PASSIVE	60 MIN-ACTIVE	AUTOMATION RELIABILITY/CONFIG
Allocation	20	20	100% Reliable (High Automation)
Router	20	20	80% Reliable (Mgt by Consent)
Image Analysis: Count Diamonds	60 (TO 30 s)	60 (TO 30 s)	80% Reliable (Mgt by Consent & Exception)
Weapon Release (Tank ATR)	60 (TO 20 s)	60 (TO 20 s)	80% Reliable (Mgt by Consent & Exception)
Change Detection	24 (TO 10 s)	24 (TO 10 s)	N/A
Qts in Chat	10	80	N/A
Systems Status	30 (TO 15 s)	240 (TO 15 s)	N/A
Digit Pairs	10 (TO 10s)	80 (TO 10s)	N/A
Comm Auditory Monitoring	32, 1 Hit/8m (TO 15 s)	240, 1 Hit/8m (TO 15 s)	N/A
Monitor Chat	20 ‘noise’	180 ‘noise’	N/A

Note. TO = time out.

The experiment used a 2 (workload: low versus high) x 2 (LOA: management-by-consent versus management-by-exception for the two surveillance tasks) mixed factorial design with workload being the between-subject variable (see Table 3).

Table 3. Experimental Design Employed in Study 1.

Subjects	Workload	Level of Automation	Condition
n = 25	Low	Management by Consent	1
		Management by Exception	2
n = 26	High	Management by Consent	3
		Management by Exception	4

The results confirmed that the high workload scenario elevated both NASA-TLX workload scores and DSSQ distress scores, compared to scores with the low workload scenario. Performance on several tasks was degraded in the high workload scenario (see Figure 7 for accuracy and neglect measures). The task load effects were stronger for the more demanding weapon release task compared to the image analysis task. Participants generally showed an appropriate level of trust (reliance) in the automation (i.e., high but not total). However, there was some variance in reliance

according to task parameters. More demanding conditions (weapon release task, high workload scenario) tended to reduce reliance, although it is in demanding conditions that trust in the automation may be most important (see Figure 8; reliance is mean percent of responses that followed the automation's recommendation). The lower level of automation (management by consent) was also associated with less reliance. However, in general, task workload level had more effect on performance than LOA did (Lin et al., 2015).

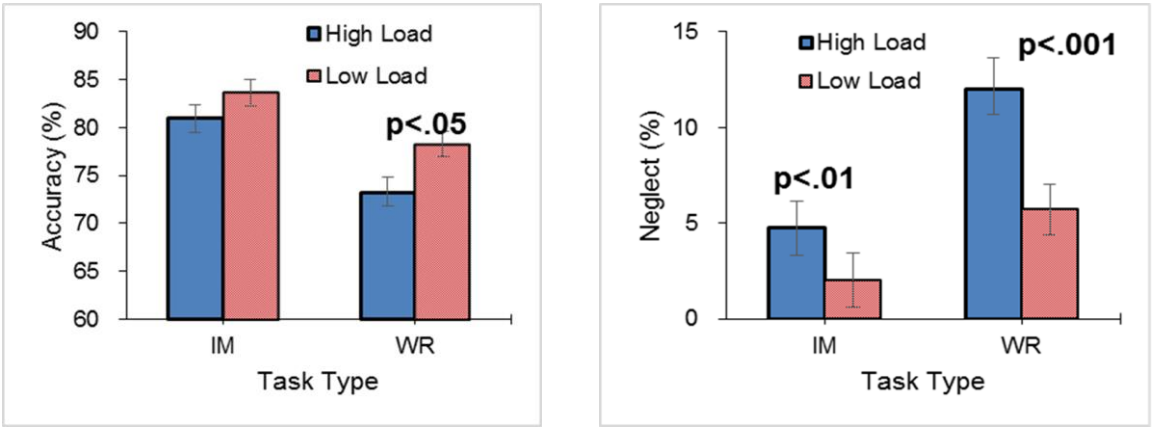


Figure 7. Mean percentage task accuracy (left) and neglect (right) for image analysis and weapon release authorization tasks as a function of task load. Bars represent standard errors.

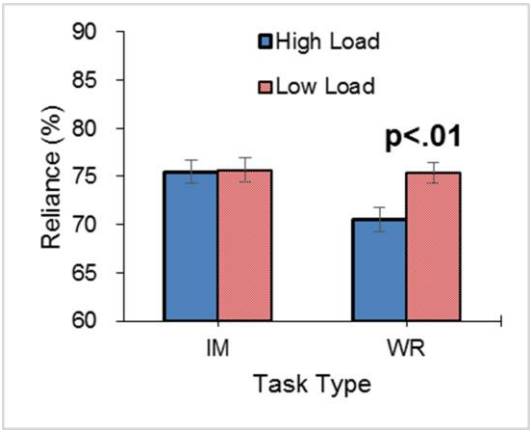


Figure 8. Mean reliance for image analysis and weapon release. Bars represent standard errors.

In terms of the individual difference data, stress state appeared to be unrelated to reliance. It was, however, associated with poorer surveillance task performance (e.g., higher neglect), as shown by high DSSQ distress and low task engagement. Results also suggested that personality predicts reliance and neglect, but not task accuracy. Participants that were higher in conscientiousness and agreeableness were less likely to neglect opening the surveillance tasks. These two personality constructs were also related to lower distress, whereas neuroticism was related to higher distress.

Video gaming expertise had more predictive value than the number of hours spent playing video games per week. Specifically, expertise was correlated with performance on the more demanding weapon release surveillance tasks. Those with higher levels of expertise, especially on action and First Person Shooter games, were more accurate and showed less task neglect. Video

gamers also relied more on the automation in demanding conditions and exhibited higher subjective task engagement and lower distress and worry. Gender differences were not statistically significant when video gaming experience was controlled.

8.2 Study 2. Fatigue and Reliability Effects

This follow-up study examined whether automation usage under sustained fatigue is moderated by the automation's reliability. More detail is available in Lin et al. (2016) and Wohleber et al. (2016). A between-subjects design was employed whereby each of the 131 participants was assigned to one of the two reliability conditions. The automation's response recommendation accuracy for both surveillance tasks (image analysis and weapons release authorization) was either 86.7% (high, *n*=67) or 60% (low, *n*=64). These reliability levels are consistent with previous studies of automation reliability (Parasuraman, Molloy, & Singh, 1993). Participants were not told the specific reliability level, just either that the automation was "quite reliable, but not perfect" (high level) or "somewhat reliable, but prone to error (low level). For this study, the experimental trial was lengthened to 120-minutes and was designed to induce passive fatigue (see Table 4 for task frequency; a trial to induce active fatigue was not utilized). For some measures, data were compared across the eight 15-minute blocks.

Table 4. Task Information including Task Frequency in 120-minute Scenario Trial.

TASK	120 MIN-PASSIVE	AUTOMATION RELIABILITY/CONFIG
Allocation	20	100% Reliable (High Automation)
Router	20	100% Reliable (High Automation)
Image Analysis: Count Diamonds	60 (TO 30 s)	60% or 86.7% Reliable (Mgt by Consent & Exception)
Weapon Release (Tank ATR)	60 (TO 20 s)	60% or 86.7% Reliable (Mgt by Consent & Exception)
Change Detection	48 (TO 10 s)	N/A
Qts in Chat	8	N/A
Systems Status	16 (TO 15 s)	N/A
Digit Pairs	16 (TO 10s)	N/A
Comm Auditory Monitoring	16, 1 Hit/30m (TO 15 s)	N/A
Monitor Chat	16 'noise'	N/A

Results showed that the automation reliability manipulation was successful. The data indicated that participants had slightly higher task accuracy with the high reliability condition compared to the low reliability condition, as well as lower stress and more subjective trust (e.g., on the HCT). However, reliance on automation differed substantially based on automation reliability level (see Figure 9). Participants relied considerably more on automation in the high reliability condition, compared to the low condition. Additionally, reliance on automation was more consistent over time with high reliability automation compared to low reliability automation. For the weapon release task, reliance on low reliability automation was more erratic and showed a

more pronounced decline than did reliance on the high reliability automation. Optimally, the reliance metric should match the objective level of reliability of the automation. However, while participants tended to be initially over-reliant on automation, relative to its reliability, by the end of the task, participants were under-reliant for weapon release. Thus, fatigue may be associated with under-utilization of automation rather than complacency, possibly because managing automation is perceived as an additional task that is shed as fatigue progresses.

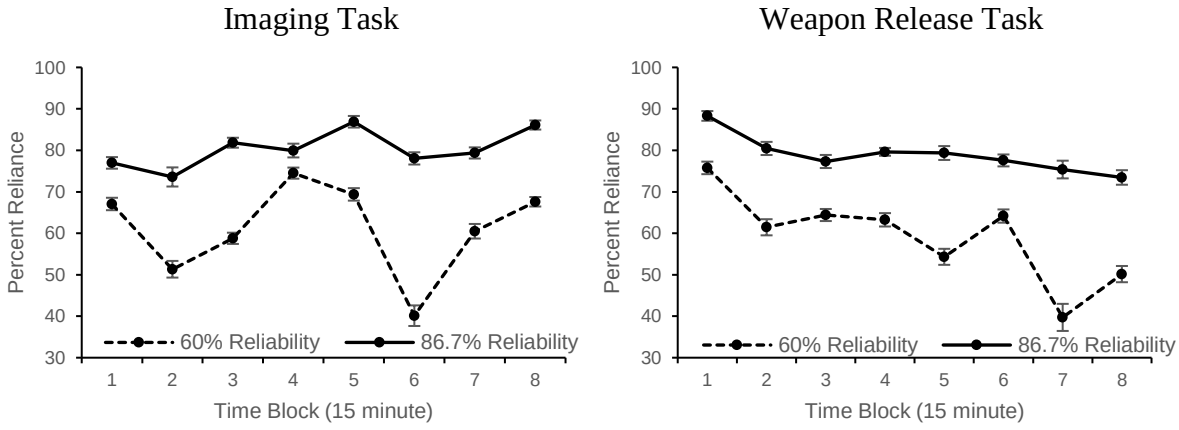


Figure 9. Percent reliance on automation in imaging analysis (left plot) and weapon release authorization (right plot) tasks as a function of reliability level.

Correlational analyses of individual difference data also showed interesting findings. Participants with higher experience and self-rated expertise in video gaming tended to be less distressed and more engaged initially, but only first person shooter game involvement predicted post-task success. Although it was expected that video gaming experience would bolster performance in the low reliability condition, no significant relationship between video gaming expertise and performance in any condition was observed. Expertise may be especially important when workload is particularly high, as in Lin et al.'s (2015) study that found benefits to gaming expertise. Inspection of the data confirmed that video gamers were less likely to rely on automation when the reliability was low, which may reflect greater performance skills. These results are consistent with those of previous studies (Abich, Matthews, & Reinerman-Jones, 2015; Lin et al., 2015) that suggest the value of selecting operators who are expert video gamers. Distress was also found associated with lower accuracy in Weapons Release, similar to Study 1, but only when reliability was low.

Analysis of the personality data showed a positive relationship between extraversion and reliance for women in the high reliability condition, and between extraversion and low accuracy for men in the low reliability condition. The trend in the data for the men is consistent with the general tendency for extraverts to perform poorly on tasks requiring sustained attention (Finomore Matthews, Shaw, & Warm, 2009). Female extraverts may compensate for this performance vulnerability by relying more on the automation when viewed trustworthy.

Gender differences were evident elsewhere in the data. The openness personality factor for men was positively correlated with reliance on high reliability automation, and was also related to conscientiousness and performance based outcomes. Particularly interesting is that the relationship between conscientiousness and mean performance was the opposite for women and men. In the low reliability condition, conscientiousness was related to higher performance and lower reliance

on automation in women, but lower performance and marginally higher reliance on low reliability automation in men. It is suspected that these results may reflect social skills (developed earlier in women and correlated to work performance; Witt & Ferris, 2003; Bennett, Farrington, & Huesmann, 2005) being related to the use of automated aids.

In regard to fatigue, the results showed a large-magnitude decline in task engagement from pre- to post-task DSSQ results, supporting the hypothesis that the simulated 120-minute RPA mission induced substantial passive fatigue (see Figure 10). The data suggest that fatigue effects may be mitigated in resilient operators, i.e., those with low distress and high task engagement. Also, if passive fatigue induced an energy conservation strategy (Sauer et al., 2003), the lack of a systematic temporal change for performance on the Image Analysis task might be explained if participants found it less effortful than completing the more difficult Weapon Release Task. Generally, participants tended to disuse the automation more with time on task as fatigue increased, especially in the low reliability condition. In that reliance patterns with fatigue were similar for both automation reliability levels, interventions developed to promote reliance optimization should be effective for both low and high reliability automation.

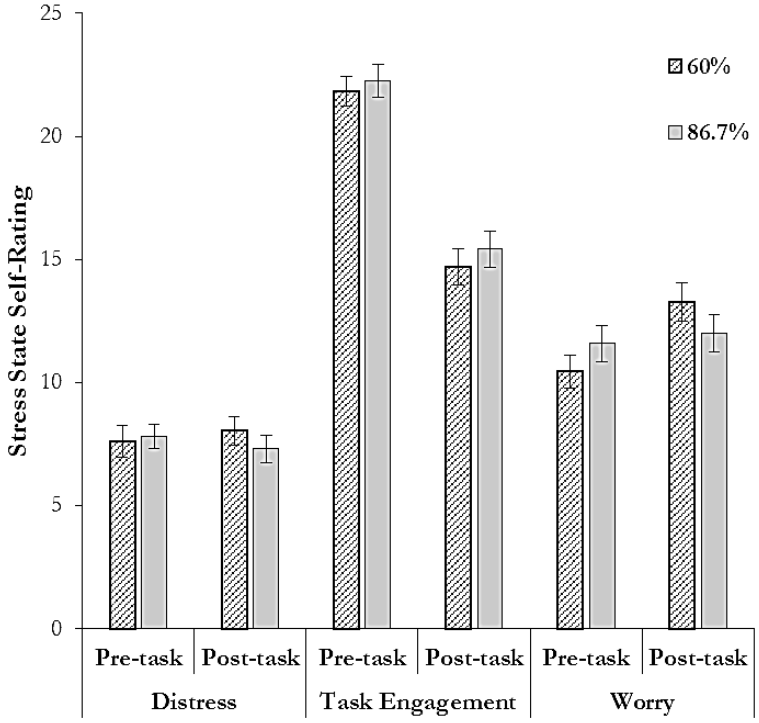


Figure 10. DSSQ distress, task engagement, and worry self-ratings pre- and post-task.

Study 2 marked the beginning of the recording of eye parameters (see Section 7.5). It was originally planned to explore the use of eye tracking metrics in examining trust in automation based on the assumption that the frequency and duration of scanning could be interpreted as indirect indicators of trust in an automated system's performance. For example, Parasuraman et al. (1993; also Parasuraman & Riley, 1997), posited that automation-induced complacency may result in less frequent scanning of automated task components. Subsequent experimentation utilizing eye tracking has generally validated this supposition (e.g., Flemisch & Onken, 2000), though automation reliability has also been demonstrated to moderate this effect (e.g., Wickens et al.,

2005). In sum, past research suggests that operator over reliance (and potentially, under reliance) on automation may be diagnosed using eye tracking metrics such as frequency and duration of fixations. In addition, eye-tracking research examining internet search behavior (e.g., Galesic et al., 2008; Guan & Cutrell, 2007) provided a method for diagnosing operator compliance with automation. This research indicated that operators typically accept the first entry in a list of automation-provided action alternatives – frequently without fully interrogating all options, to arrive at a rational, informed decision. This suggests that metrics such as serial position and average dwell time during operator interaction with automation-provided actions may index operator compliance.

Finally, there is some possibility that operator fatigue may interact with automation induced complacency, such that fatigued operators may be more likely to rely on automation (i.e., by more frequent failures to detect automation 'misses') and comply with automation (i.e., by accepting inappropriate automation recommendations more frequently) than when they are not fatigued. This outcome would be demonstrated in metrics of eye gaze behavior by operators fixating less frequently and for shorter durations on task aspects that are automated, and by operators spending less time fixated on automation-recommended courses of action (as an index of participants' evaluation of those options).

Unfortunately in the present effort, the integration of the eye tracker with the ALOA simulation was problematic, limiting the degree to which eye parameters could be applied in the manner accomplished in the research reviewed above. By applying the faceLAB quality scale (range 0-3), only a subset of participants (N = 39) met our criteria of having over 75% of the samples having a score of at least 2 (one eye tracking) or 3 (full tracking). One issue was the difficulty in maintaining track of the eye with the ALOA system because it employed two 24-in widescreen monitors. It was still difficult to maintain track of participants' gaze point (study 3 switched to two 20-in standard monitors at 1680 x 1050 resolution with marked improvement in tracking quality). Moreover, the sizes of the elements within the task windows that might inform automation usage (e.g., fixating on one response option versus another or whether the participant skipped analyzing the image and just looked and accepted the automation's recommendation) were too small to allow for confident fixation target discrimination with the current apparatus.

The results from this subset of participants showed that the ocular measures were typically more sensitive to time on task than to the level of automation reliability (e.g., see Figure 11).

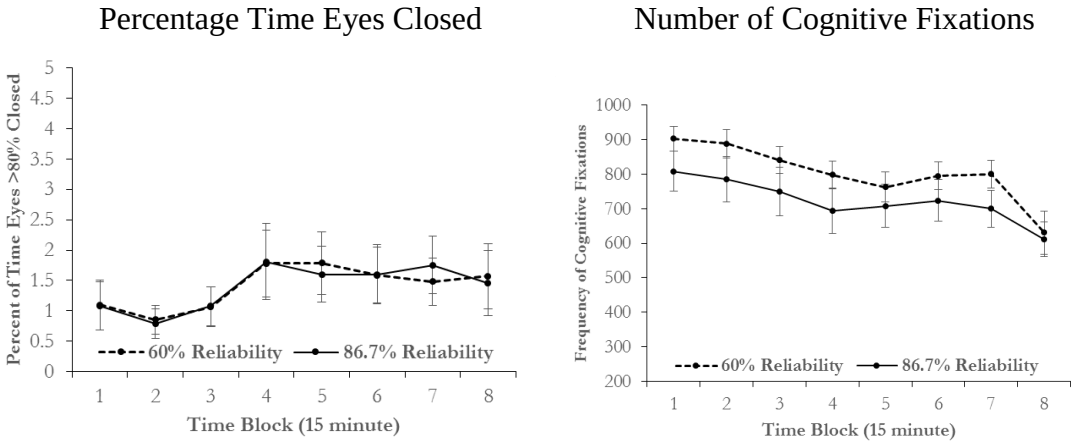


Figure 11. Percentage of time eyes >80% closed and frequency of cognitive fixations for low and high automation reliability levels as a function of time block.

An additional analysis was conducted examining the mean percentage of time participants spent gazing at each element of the of ALOA display. To accomplish this analysis, the display was partitioned into separate areas of interest (AOI's) where each represented the separate sub-task elements of the ALOA simulation (shown in Figure 1). However, division of some areas was not feasible because the sub-task display elements were essentially too small for the eye tracker to reliably distinguish, leading to the combination of some task elements into aggregate AOI's. This approach resulted in seven distinct AOI's presented in Figure 12 below. AOI 1, which corresponds with the image tasks, was further subdivided into two regions reflecting the image and response elements of the task.

A) Left ALOA monitor

B) Right ALOA monitor

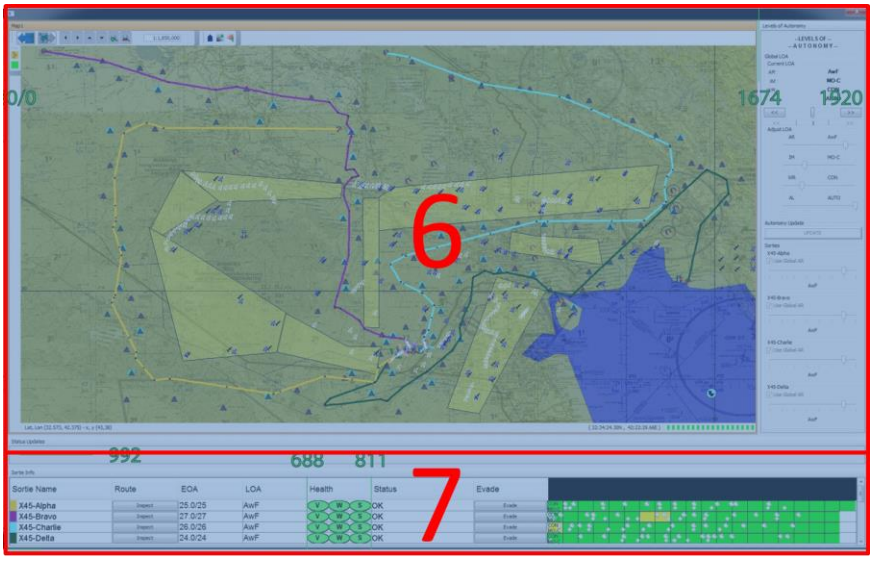
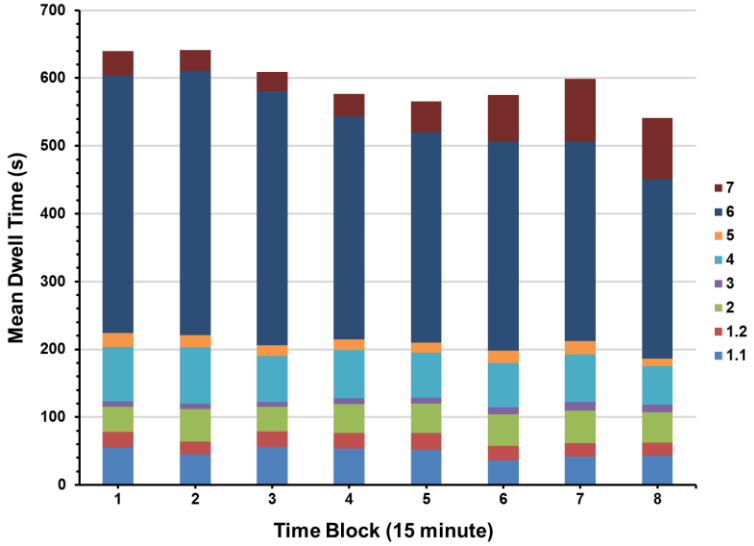


Figure 12. ALOA areas of interest designated for dwell time analysis. Panel A and B reflect the ALOA sub-tasks presented to participants on the left and right monitors, respectively.

Analysis of the cumulative time participants spent gazing at each AOI (i.e., dwell time) was then conducted utilizing a 2 (reliability) x 8 (time block) x 8 (AOI) mixed ANOVA. The results of this analysis revealed statistically significant main effects for time block, F (4.13, 156.81) = 13.66, $p < .05$, $\eta_p^2 = .264$, and AOI, F (1.42, 53.98) = 369.87, $p < .05$, $\eta_p^2 = .907$, a statistically significant time block by AOI interaction, F (7.03, 267.04) = 30.61, $p < .05$. $\eta_p^2 = .446$, and a statistically significant reliability by time block by AOI interaction, F (7.03, 267.04) = 2.19, $p < .05$, $\eta_p^2 = .054$. No other sources of variance in the analysis were statistically significant (all $p > .05$). As can be seen in Figure 13, participants spent the most time gazing at AOI 6 (the main RPA display), and that, overall, they reduced the amount of time they spent gazing at the display across time blocks.

A) Low reliability



B) High reliability

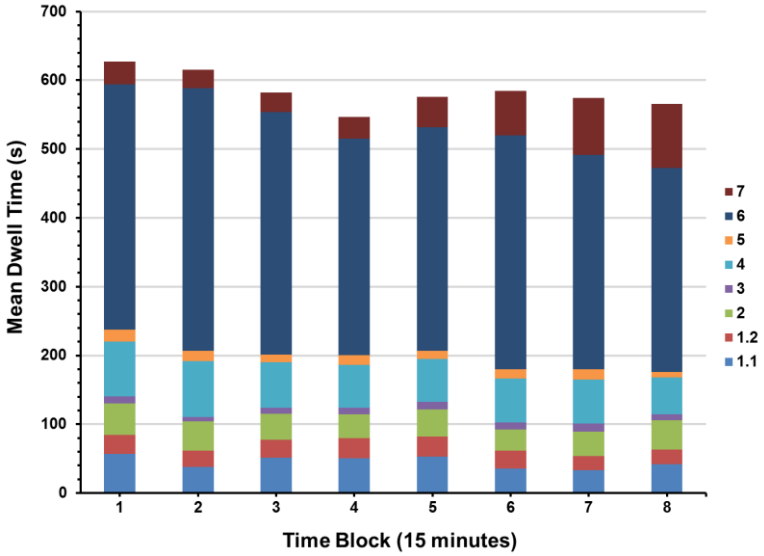


Figure 13. Mean dwell time (seconds) in each AOI per time block for the low (Panel A) and high (Panel B) automation reliability conditions.

To further explicate the reliability by time block by AOI interaction, follow up post hoc simple effects analyses were conducted separately examining the effects of reliability and period at each AOI. The results of these analyses indicated that mean dwell time changed as a function of period in AOI's 1.1, 1.2, 4, 5, 6, and 7, but not in AOI's 2 and 3 (both aspects of the allocation task). However, the observed changes in dwell time across AOI's were best described by complex polynomials (order 5-7), suggesting that changes cannot simply be ascribed to time-on-task effects. With regard to reliability condition, the simple effects analyses indicated it had little effect on dwell time (i.e., there were no statistically significant main effects of reliability, nor any statistically significant reliability by time block interactions, all $p > .05$).

For a more complete accounting of eye tracking metrics and outcomes, please see Wohleber (2016).

8.3 Nonlinear Analyses of Eye Gaze Behavior

As part of this program of research, several alternative eye gaze analysis approaches were explored for detection of fatigue and operator reliance. These approaches primarily utilized nonlinear time series analysis, such as recurrence quantification analysis (RQA), to characterize participants' fixations for patterns, and to examine if those patterns were differentiable based on reliability condition and time-on-task. Short abstracts broadly describing each effort are presented below.

Alternative indices of performance: An exploration of eye gaze metrics in a visual puzzle task. Of interest to the U.S. Air Force is the ability to develop and characterize the level of workload that operators are under at any given point. When an operator's cognitive resources exceed demands, a 'red line' of performance may be crossed after which performance breaks down. What is needed is an estimate of operator state; a 'dipstick' for the operator in order to assess the level of 'resources' available, in order to avoid performance problems. Traditional approaches use secondary tasks (e.g., mental arithmetic) or secondary physiological measures (e.g., heart rate variability) for state assessment. However, this study was motivated by dynamic systems theory which indicates that there are meaningful patterns of variability in 'primary' behaviors (e.g., required activities) which might provide a measure of operator state. Eye gaze was utilized as a primary measure in a visual puzzle task. The link between eye gaze and attention is generally accepted as is the link between attention and performance outcomes. The goal of Experiment 1 was to determine if performance changes in a visual puzzle task were reflected in eye gaze, as measured in multiple ways, i.e., conventional (e.g., average fixation length) and dynamic (e.g., β values, measures derived from a recurrence matrix) indices. These relationships were explored in relation to task difficulty, time on task, as well as spare capacity. The results of Experiment 1 suggest that there are impacts of task demands on gaze patterns, for both conventional and dynamic gaze metrics. There were also significant effects of practice on eye gaze patterns in Experiment 1 that could be interpreted as learning or strategy shifts with time-on-task. The impact of learning on eye gaze was explored in a follow up experiment. The results of Experiment 2 show a significant improvement in performance in the task accompanied by change in gaze patterns when repeating the same puzzle; and that the dynamic measure of diagonal recurrence was systematically related to this performance change. This suggests that non-conventional measures of dynamic structure provide additional and complimentary information about operator state.

For a more complete accounting of this research, please see Russell et al. (2014).

Effects of automation reliability on structure of gaze patterns over time. Air Force operations are becoming increasingly automated, exemplified by the vision to enable single operator supervision of multiple RPAs. While increased automation is a requirement for this vision, it also introduces human factors issues. In particular, the operator may become inappropriately reliant on automation, potentially compromising human-automation system performance. One route to monitor operator reliance is gaze tracking; however, few experiments have focused on this approach. Several recent studies suggest that metrics such as the frequency and pattern of visual inspection of automated system components may provide objective indices of operator trust and reliance on automation. Specifically, the regularity of the pattern of transitions between areas of interest was evaluated as well as the homogeneity of the distribution of gaze as a function of time and reliability of automation. Hierarchical linear mixed model analysis of the regularity of transitions between areas of interest indicated a significant difference between high and low reliability conditions as a function of time. Individuals in the high reliability condition tended to linearly decrease in regularity of gaze patterns over the first half of the session, but exhibited a higher degree of second order growth than groups in the low reliability condition, who exhibited an overall higher degree of regularity in gaze patterns by the end of the session. Additionally, analysis of the Shannon entropy of the distribution of gaze duration binned into 500 by 500 pixels showed that individuals in the high reliability condition tended to become more focused on particular regions of the display over time, whereas groups in the low reliability condition did not. Together, these results indicate the high reliability automation increased the regularity of gaze patterns of individuals over low reliability conditions, perhaps by allowing individuals to focus on key areas rather than distributing their attention more uniformly.

For a more complete accounting of this research, please see Tolston et al. (2016).

8.4 Study 3. Diagnostic Fatigue Monitoring for Adapting Level of Automation

A further advantage of eye tracking, in the context of the current proposal, is that substantial research has been conducted linking changes in eye gaze behavior to states of fatigue. Perhaps the most robust effects reported in the literature are that states of fatigue increase blink rate and blink duration (see e.g., Stern, Boyer, & Schroeder, 1994, and Schleicher et al., 2008, for reviews of the relevant literature). In addition, the percentage of eye closure (PERCLOS; Wierwille et al., 1994) has been identified as a useful indicator (e.g., Dinges, Mallis, Maislin, & Powell, 1998; McKinley, McIntire, Schmidt, Repperger, & Caldwell, 2011), though some recent research suggests that it is influenced by individual differences (e.g., Schleicher et al., 2008; van Orden, Jung, & Makeig, 2000), potentially reducing its overall utility as a fatigue indicator, unless individual differences may be factored into the diagnostic model.

Building on the results from Study 2, a study is in progress to explore how real-time eye parameter recordings might be employed to diagnose fatigue state and, in turn, drive interface adaptations and improve task performance. Its objective is to examine the utility of an ocular-based fatigue detecting algorithm for adapting task LOA. The thresholds employed in the algorithm are based on the analysis of ocular data collected in Study 2. Specifically, the frequency of express (less than 150 ms) and cognitive (150-900 ms) fixations for each 15-minute periods of a 120-minute scenario will be determined. Once the pre-determined threshold for triggering fatigue mitigation has been met, the LOA of the two surveillance task will be reduced for the next 15-minute period from one requiring a consent response to one that has no automated aiding. The

LOA will then revert back during the subsequent period. Success in meeting this objective will be demonstrated by participants in the adaptive automation condition outperforming participants assigned to a condition with static LOA.

9. CONCLUSIONS

The effective application of automation to future systems is a significant priority of the Air Force (Dahm, 2010; Endsley, 2015). While manned cockpits have become increasingly automated, the level of system automation present in most RPAs is even greater. The need for future pilots to supervise multiple RPAs adds to the critical role of automation (Eggers & Draper, 2006). Existing psychological research on automated systems has established beneficial effects of automation on operator workload and situation awareness, as well as identifying a range of threats to operational effectiveness (Parasuraman & Mouloua, 1996). Both over- and under-reliance on automation may jeopardize effectiveness of the human-automation system (Parasuraman & Riley, 1997). Thus, it is timely to conduct basic research on automation usage issues during sustained operations with the goal of enhancing the designs of human-automation interfaces for future RPA systems. It is especially important to support operators working in conditions of fatigue and stress that are prevalent in RPA missions (Ouma, Chappelle, & Salinas, 2011).

This research effort extends knowledge pertaining to human-automation in several respects. First, the studies demonstrate that the nature and cognitive demands of the RPA task itself may pose significant human factors challenges. Workload demands of RPA operation are known to be highly variable. Here, the high workload was associated with time pressure and multitasking-induced large-magnitude increases in distress (Study 1), whereas low workload and a monotonous, longer-duration mission produced passive fatigue and loss of task engagement (Study 2). As expected, performance deficits were also obtained associated with higher workload (Study 1) and lower automation reliability (Study 2). In addition, Study 2 suggested a vigilance-like temporal decrement in accuracy on the more demanding Weapon Release task, especially when automation reliability was low.

Although participants were appropriately sensitive to automation reliability (Study 2), there were several instances in which reliance on automation was suboptimal. In Study 1, participants tended to under-rely on automation, especially for Weapons Release and when the LOA was set to management-by-consent. Thus, participants do not turn to assistance from the automation in the higher demand conditions in which reliance would be appropriate. In Study 2, participants became increasingly under-reliant on automation for Weapons Release over time, despite a tendency towards deteriorating performance over time for this ISR task component. Generally, participants do not seem adept at using automation strategically to compensate for performance shortcomings, which would be a concern in any practical application.

Second, study findings identified several individual difference factors that may influence operator performance and reliance on automation. Individual differences in stress response were associated with performance on ISR tasks, consistent with previous studies (Matthews et al., 2013). Generally, high distress and low task engagement were associated with less accurate performance and neglect of mission objectives. However, these associations were more pronounced when the task was configured to be more demanding due to high workload (Study 1) or unreliable automation (Study 2). Stress states were not consistently associated with reliance on automation. Instead, several personality factors were discovered that predicted reliance, although correlations varied with task factors, similar to previous studies of the role of personality in

automated systems (Szalma & Taylor, 2011). Results showed that self-rated expertise in video gaming, especially with action games, was associated with both more effective performance, and better trust calibration, but only when workload was high (Study 1). Cognitive-attentional skills acquire through action gaming may enhance multi-tasking under time pressure. Video gamers also showed higher levels of subjective task engagement in both studies, suggesting some general resilience to task stress.

Third, findings suggest strategies for mitigating performance deficits and suboptimal use of automation. Correlational findings suggest that selection of stress-resilient operators may generally benefit mission outcomes, although these individuals have no special advantage in automation use. Recruitment of experienced action gamers would also be advantageous, especially for high-workload missions. Findings also suggested that, with video gaming experience controlled, female operators may perform as well as male ones. However, predictors of operator performance and trust may vary somewhat by gender. A second mitigation strategy is to develop finely-tuned training interventions to target specific vulnerabilities. Given that suboptimal use of automation and performance deficits depended on task configuration, training interventions might be developed for problematic configurations. For example, participants might be trained to rely more on automation to mitigate high workload. At the least, training should incorporate scenarios varying in workload, level of automation, and automation reliability. The role of personality factors in reliance suggests that training could also be tailored to individual characteristics, although clarification of the seemingly complex role of personality traits in trust may be necessary. A third form of intervention is to use diagnostic monitoring of operator neurocognitive status to drive adaptive automation. Data suggest that eye tracking may provide a nonintrusive means to operator monitoring, although there are practical issues surrounding tracking across dual screens, and it appears to be easier to monitor for fatigue than for reliance optimization. Study 3 will provide further evidence on the feasibility of the approach. Work is underway to implement eye tracking metrics measures into an adaptive interface designed to adjust LOA to levels of operator fatigue.

Taken together, these results provide a better understanding of the circumstances under which individual differences (e.g., video game experience), fatigue, and automation characteristics may interact to produce inappropriate reliance on automation. They demonstrate that in conducting research and evaluating RPA automation, it is critical to examine effects across a range of task configurations. In addition, the utility of eye tracking to diagnose suboptimal use of automation was explored. While single operator supervisory control of multi-RPAs was employed as the problem domain to support the proposed research, these findings should inform human-automation interaction in several domains including vehicle operation, process control, and medicine. Future applications for the outcomes of this research also include selection and training, operator performance assessment, and online adaptive aiding.

10. REFERENCES

Abich, J., Matthews, G., & Reinerman-Jones, L. E. (2015). Individual differences in UGV operation: A comparison of subjective and psychophysiological predictors. *Proceedings of the Human Factors and Ergonomics Society, 59*, 741-745.

Anderson, C. A., Gentile, D. A., & Buckley, K. (2007). *Violent video game effects on children and adolescents: Theory, research, and public policy*. New York: Oxford University Press.

Bennett, S., Farrington, D. P., & Huesmann, L. R. (2005). Explaining gender differences in crime and violence: The importance of social cognitive skills. *Aggression and Violent Behavior, 10,* 263-288.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America, 107,* 1065-1066.

Bolia, R. S., Nelson, W. T., Middendorf, M. S., Guilliams, N. M., & McLaughlin, A. B. (2004). Evaluating the utility of a multi-layer visual display for air battle managers. *Proceedings of the Human Factors and Ergonomics Society, 48,* 21-25.

Burkolter, D., Kluge, A., & Matthias, B. (2010). Individual differences in complex task performance: Interaction effects of risk-taking behavior and cognitive variables. *Proceedings of the Human Factors Society, 54,* 2333-2337.

Calhoun, G. L., Ruff, H. A., Draper, M. H., & Wright, E. J. (2011). Automation-level transference effects in simulated multiple unmanned aerial vehicle control. *Journal of Cognitive Engineering & Decision Making, 5,* 55-82.

Calhoun, G. L., Ruff, H. A., & Murray, C. (2012). Multi-unmanned vehicle supervisory control: An initial evaluation of personality drivers. *AIAA Infotech@ Aerospace Conference,* 1-22.

Chappelle, W., McDonald., K., & King, R. E. (2010*). Psychological attributes critical to the performance of MQ-1 Predator and MQ-9 Reaper U.S. Air Force sensor operators.* Technical Report AFRL-SA-BR-TR-2010-0007, USAF School of Aerospace Medicine, Brooks City-Base, TX.

Chen, J. Y. C., & Barnes, M. J. (2012). Supervisory control of multiple robots: Effects of imperfect automation and individual differences. *Human Factors, 54,* 157-174.

Chen, J. Y. C., Procci, K., Boyce, M. W., Wright, J. L., Garcia, A., & Barnes, M. J. (2014). *Situation Awareness-based Agent Transparency* (Final No. ARL-TR-6905). Aberdeen Proving Ground, MD: Army Research Lab.

Chen, J. Y. C. & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics, 52,* 907-920.

Cummings, M. L., Clare, A., & Hart, C. (2010). The role of human-automation consensus in multiple unmanned vehicle scheduling. *Human Factors, 51*(1), 17-27.

Dahm, W. (2010). *Chief scientist's report on technology horizons: A vision for Air Force science and technology during 2010-2030.* Volume 1, Air Force Technical Report, AF/ST-TR-10-01.

de Greef, T., Lafeber, H., van Oostendorp, H., & Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. In D. D. Schmorrow et al. (Eds.), *Lecture notes in computer science: Vol. 5638. Foundations of augmented cognition: Neuroergonomics and operational neuroscience* (pp. 219-228). Berlin, Germany: Springer-Verlag.

Desmond, P. A., & Hancock, P. A. (2001). Active and passive fatigue states. In P.A. Hancock & P.A. Desmond (Eds.), *Stress, workload, and fatigue* (pp. 455-465). Mahwah, NJ: Lawrence Erlbaum.

Dinges, D. F., Mallis, M. M., Maislin, G., & Powell, J. W. (1998). *Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management* (Report No. DOT HS 808 762). Washington, DC: National Highway Traffic Safety Administration.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse the misses? *Human Factors, 49,* 564-572.

Donmez, B. D., Nehme, C., & Cummings, M. L. (2010). Modeling workload impact in multiple unmanned vehicle supervisory control. *IEEE Systems, Man, and Cybernetics, Part A Systems and Humans, 40,* 1180-1190.

Eggers, J. W., & Draper, M. H. (2006). Multi-UAV control for tactical reconnaissance and close air support missions: Operator perspectives and design challenges. *Proceedings of the NATO RTO Human Factors and Medicine Panel Symposium HFM-135* held in Biarritz, France, 9-11 Oct. NATO RTO: Neuilly-sur-Siene, CEDEX. Biarritz, France.

Endsley, M. (2015). *Autonomous horizons: System autonomy in the Air Force – A path to the future.* Washington, DC: Office of the Chief Scientist.

Fidopiastis, C. M., Drexler, J., Barber, D., Cosenzo, K., Barnes, M., Chen, J. Y. C., & Nicholson, D. (2009). Impact of automation and task load on unmanned system operator's eye movement patterns. In D. D. Schmorrow et al. (Eds.), *Lecture Notes in Computer Science: Vol. 5638. Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience* (pp. 229-238). Berlin, Germany: Springer-Verlag.

Finomore, V., Matthews, G., Shaw, T., & Warm, J. (2009). Predicting vigilance: A fresh look at an old problem. *Ergonomics, 52,* 791-808.

Flemisch, F. O., & Onken, R. (2000). *Detecting usability problems with eye tracking in airborne battle management support* (Report ADPO10701). Washington, DC: Defense Technical Information Center.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: new insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72,* 892-913.

Gentile, D. (2009). Pathological video-game use among youth ages 8 to 18: A national study. *Psychological Science, 20,* 594-602.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26-42.

Green, C. S., & Bavelier, D. (2008). Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychology and Aging, 23,* 692-701.

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 417–420). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1240691

Guznov, S. Y. (2011). *Visual search training techniques in a simulated UAV environment.* Doctoral dissertation, University of Cincinnati.

Guznov, S., Matthews, G., Funke, G., & Dukes, A. (2011). Use of the RoboFlag synthetic task environment to investigate workload and stress responses in UAV operation. *Behavior Research Methods, 43,* 771-780.

Guznov, S., Matthews, G., & Warm, J. S. (2010). Team member personality, performance, and stress in a Roboflag synthetic task environment. *Proceedings of the Human Factors and Ergonomics Society, 54,* 1679-1683.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 239-250). Amsterdam: North-Holland.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *International Journal of Aviation Psychology, 9*, 19-32.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology, 45*, 73-93.

Hockey, G. R. J., Wastell, D. G., & Sauer, J. (1998). Effects of sleep deprivation and user interface on complex performance: A multilevel analysis of compensatory control. *Human Factors*, *40*, 233-253.

Johnson, R., Leen, M., & Goldberg, D. (2007). *Testing adaptive levels of automation (ALOA) for UAV supervisory control*. Whittier, CA: OR Concepts Applied.

Kidwell, B., Calhoun, G., Ruff, H., Parasuraman, R. (2012). Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles. *Proceedings of the Human Factors and Ergonomics Society*, *56*, 428-432.

Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for automation. *Human Factors, 46*, 50-80.

Lin, J., Matthews, G., Wohleber, R. W., Chiu, C.-Y. P., Calhoun, G. L., Funke, G. J., & Ruff, H. A. (2016). Automation reliability and other contextual factors in multi-UAV operator selection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*.

Lin, J., Wohleber, R. W., Matthews, G., Chiu, C.-Y. P., Calhoun, G. L., Ruff, H. A., Funke, G. J. (2015). Videogame experiences and gender as predictors of performance and stress during supervisory control of multiple unmanned aerial vehicles. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*, *59*, 746-750.

Liu, D., Wasson, R., & Vincenzi, D. A. (2009). Effects of system automation management strategies and multi-mission operator-to-vehicle ratio on operator performance in UAV systems. *Journal of Intelligent and Robotic Systems*, *54*, 795–810.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of Eleventh Australasian Conference on Information Systems* (pp. 6–8). Citeseer.

Matthews, G., Campbell, S.E., Falconer, S., Joyner, L., Huggins, J., Gilliland, K., Grier, R., & Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress and worry. *Emotion, 2*, 315-340.

Matthews, G., Davies, D. R., Westerman, S. J., & Stammers, R. B. (2000). *Human performance*: *Cognition, stress and individual differences*. London: Psychology Press.

Matthews, G., & Desmond, P. A. (2002). Task-induced fatigue states and simulated driving performance. *Quarterly Journal of Experimental Psychology, 55A*, 659-686.

Matthews, G., Desmond, P. A., & Hitchcock, E. M. (2012). Dimensional models of fatigue. In G. Matthews, P. A. Desmond, C. Neubauer & P. A. Hancock (Eds.), *Handbook of operator fatigue* (pp. 139-154). Aldershot, UK: Ashgate Press.

Matthews, G., Hancock, P. A., & Desmond, P. A. (2012). Models of individual differences in fatigue for performance research. In G. Matthews, P.A. Desmond, C. Neubauer & P.A. Hancock (Eds.), *Handbook of operator fatigue* (pp. 155-170). Aldershot, UK: Ashgate Press.

Matthews, G., Neubauer, C. E., Saxby, D. J., & Langheim, L. K. (2012). Driver fatigue: The perils of vehicle automation. In M. Sullman & L. Dorn (Eds.), *Proceedings of the International Congress of Applied Psychology (Traffic and Transportation)* (pp. 127-139). Ashgate, UK: Aldershot.

Matthews, G., Szalma, J., Panganiban, A.R., Neubauer, C., & Warm, J.S. (2013). Profiling task stress with the Dundee Stress State Questionnaire. In L. Cavalcanti & S. Azevedo (Eds.), *Psychology of stress: New research* (pp. 49-90). Hauppage, NY: Nova Science.

Matthews, G., Warm, J. S., Reinerman, L. E., Langheim, L. K., & Saxby, D. J. (2010a). Task engagement, attention and executive control. In A. Gruszka, G. Matthews & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory and executive control* (pp. 205-230). New York: Springer.

Matthews, G., Warm, J. S., Reinerman, L. E., Langheim, L. K., Washburn, D.A., & Tripp, L. (2010b). Task engagement, cerebral blood flow velocity, and diagnostic monitoring for sustained attention. *Journal of Experimental Psychology: Applied, 16,* 187–203.

McKinley, R.A., McIntire, L. K., & Funke, M.A. (2011). Operator selection for unmanned aerial systems: Comparing video game players and pilots. *Aviation, Space and Environmental Medicine, 82,* 635-642.

McKinley, R.A., McIntire, L. K., Schmidt, R., Repperger, D. W., & Caldwell, J.A. (2011). Evaluation of eye metrics as a detector of fatigue. *Human Factors, 53,* 403-414.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interaction. *Human Factors, 50,* 194-210.

Mikulka, P. J., Scerbo, M. W., & Freeman, F. G. (2002). Effects of a biocybernetic system on vigilance performance. *Human Factors, 44,* 654-664.

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: delegation interfaces for supervisory control. *Human Factors, 49,* 57-75.

Mouloua, M., Gilson, R., & Hancock, P. (2003). Human-centered design of unmanned aerial vehicles. *Ergonomics in Design, 11,* 6-11.

Neubauer, C., Matthews, G., Langheim, L., & Saxby, D. (2012). Fatigue and voluntary utilization of automation in simulated driving. *Human Factors, 54,* 734–746.

Ouma, J., Chappelle, W., & Salinas, A. (2011). *Facets of occupational burnout among U.S. Air Force active duty and National Guard/Reserve MQ- 1 Predator and MQ-9 Reaper operators*. Technical Report, AFRL-SA-WP-TR-2011-0003, Air Force Research Laboratory, 711th Human Performance Wing, School of Aerospace Medicine, Wright-Patterson AFB, OH.

Parasuraman, R., & Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology, 3,* 1-23.

Parasuraman, R., & Mouloua, M. (1996). *Automation and human performance: Theory and applications*. Mahway, NJ: Erlbaum.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-253.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 50,* 511-520.

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology, 40,* 187-195.

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. *Human Factors, 45,* 601-613.

Rice, S., & Keller, D. (2009). Automation reliance under time pressure. *International Journal of Cognitive Technology, 14,* 36-44.

Richardson, A. E., Powers, M. E., & Bousquet, L. G. (2011). Video game experience predicts virtual, but not real navigation performance. *Computers in Human Behavior, 27*, 552-560.

Ruff, H.A. & Calhoun, G.L. (2011). Impact of automation's level and reliability on multi-vehicle supervisory control task performance. *AIAA Infotech @ Aerospace Conference, AIAA-2011-1489-102.*

Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence: Teleoperators and Virtual Environments, 11*, 335-351.

Russell, S.M., Funke, G.J., Miller, B.T., Dukes, A., Flach, J.M., Watamaniuk, S.N.J., Strang, A.J., Menke, L., & Brown, R. (2014). Alternative indices of performance: An exploration of eye gaze metrics in a visual puzzle task (Technical Report AFRL-RH-WP-TR-2014-0095). Air Force Research Laboratory, 711 Human Performance Wing, Human Effectiveness Directorate, Wright-Patterson AFB.

Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality, 36,* 1-31.

Sauer, J., Wastell, D. G., Hockey, G. R. J., & Earle, F. (2003). Performance in a complex multiple-task environment during a laboratory-based simulation of occasional night work. *Human Factors, 45*, 657-669.

Saxby, D. J., Matthews, G., Hitchcock, E., & Warm, J. S. (2007). Development of active and passive fatigue manipulations using a driving simulator. *Proceedings of the Human Factors and Ergonomics Society, 51,* 1237-1241.

Saxby, D. J., Matthews, G., Hitchcock, E. M., Warm, J. S., Funke, G. J., & Gantzer, T. (2008). Effects of active and passive fatigue on performance using a driving simulator. *Proceedings of the Human Factors and Ergonomics Society, 52,* 1751-1755.

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics, 51*, 982-1010.

Schulte, A., & Donath, D. (2011). Measuring self-adaptive UAV operators' load-shedding strategies under high workload. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics, HCI 2011, LNAI 6781* (pp. 342-351). Berlin: Springer-Verlag.

Shaw, T. H., Matthews, G., Warm, J. S., Finomore, V., Silverman, L., & Costa, P. T. (2010). Individual differences in vigilance: Personality, ability and states of stress. *Journal of Research in Personality, 44,* 297-308.

Spence, I., & Feng, J. (2010). Video games and spatial cognition. *Review of General Psychology, 14*(2), 92-104.

Spence, I., Yu, J. J. J., Feng, J., & Marshman, J. (2009). Women match men when learning a spatial skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1097-1103.

Stern, J.A., Boyer, D., & Schroeder, D. (1994). Blink rate: A possible measure of fatigue. *Human Factors, 36*, 285-297.

Szalma, J. L., & Taylor, G. S., (2011). Individual differences in response to automation: The Five Factor Model of personality. *Journal of Experimental Psychology: Applied, 17,* 71-96.

Szalma, J. L., & Teo, G. W. L. (2010). The joint effect of task characteristics and neuroticism on the performance, workload, and stress of signal detection. *Proceedings of the Human Factors Ergonomics Society, 54*, 1052-1054.

Tang, Y., & Posner, M. I. (2009). Attention training and attention state training. *Trends in Cognitive Sciences, 13*, 222-227.

Tolston, M., Funke, G., Matthews, G., Wohleber, R., Lin, J., Calhoun, G., & Ruff, H. (2016, July). *Effects of automation reliability on structure of gaze patterns over time*. Poster to be presented at the 7th International Conference on Applied Human Factors and Ergonomics, Orlando, FL.

Tvaryanas, A. P., & MacPherson, G. D. (2009). Fatigue in pilots of remotely piloted aircraft before and after shift work adjustment. *Aviation, Space, and Environmental Medicine, 80,* 454-461.

van Orden, K. F., Jung, T. P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological Psychology, 52,* 221-240.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors, 50,* 433-441.

Wickens, C. D., Dixon, S. R., Goh, J., & Hammer, B. (2005, April). *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis*. Paper presented at the 13th Annual International Symposium of Aviation Psychology, Oklahoma City, OK.

Wierwille, W. W., Ellsworth, L.A., Wreggit, S. S., Fairbanks, R. J., & Kirn, C. L. (1994). *Research on vehicle based driver status/performance monitoring: Development, validation, and refinement of algorithms for detection of driver drowsiness* (Report No. DOT HS 808 247). Washington, DC: National Highway Traffic Safety Administration.

Witt, L. A., & Ferris, G. R. (2003). Social skill as moderator of the conscientiousness-performance relationship: Convergent results across four studies. *Journal of Applied Psychology, 88,* 809-820.

Wohleber, R. W. (2016). *The impact of automation reliability and fatigue on reliance* (Doctoral dissertation). University of Central Florida. Retrieved from https://www.researchgate.net/publication/303518932_The_impact_of_automation_reliability_and_fatigue_on_reliance

Wohleber, R. W., Calhoun, G. L., Funke, G. J., Ruff, H. A., Chiu, C.-Y. P., Lin, J., & Matthews, G. (2016). The impact of automation reliability on performance and reliance changes with operator fatigue. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*.

10.1 Contract Publications and Presentations

Lin, J., Matthews, G., Wohleber, R. W., Chiu, C.-Y. P., Calhoun, G. L., Funke, G. J., & Ruff, H. A. (2016). Automation reliability and other contextual factors in multi-UAV operator selection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*.

Lin, J., Wohleber, R. W., Matthews, G., Chiu, C.-Y. P., Calhoun, G. L., Ruff, H. A., Funke, G. J. (2015). Videogame experiences and gender as predictors of performance and stress during supervisory control of multiple unmanned aerial vehicles. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting, 59,* 746-750.

Matthews, G. (2014, July). Distraction, fatigue and the automated vehicle. Invited address, *International Congress for Applied Psychology*, Paris.

Matthews, G. (2015, April). Task-induced fatigue, operator engagement and the control of attention. *Human Factors in Control Forum on Training for Unexpected and Emergent Situations*, Bergen, Norway.

Matthews, G. (2015, May). Trust in autonomy Panel Session. *DoD HFE TAG Meeting 69*, Orlando, FL.

Matthews, G. (2015, July). Task-induced fatigue, operator engagement and the control of attention. Presidential address, *Seventeenth Meeting of the International Society for the Study of Individual Differences*, London, Ontario.

Matthews, G. (2016, June). Multivariate workload assessment strategies for autonomous vehicle teaming. *Air Force Blue Sky meeting on Workload Assessment in the Context of Autonomy*, Pensacola, FL.

Matthews, G. (2016, July). Fatigue and workload management in autonomous and unmanned vehicle operation. Invited address, *International Congress of Psychology*, Yokohama, Japan.

Matthews, G. (in press). Multidimensional profiling of task stress states for human factors: A brief review. *Human Factors.*

Matthews, G., Lin, J., Wohleber, R., & Reinerman-Jones, L. (2015, July). Individual differences in unmanned vehicle operation: Performance, stress and trust. *Annual Meeting of the International Society for the Study of Individual Differences*, London, ON, Canada.

Matthews, G., Lin, J., Wohleber, R., Reinerman-Jones, L., Calhoun, G., & Chiu, P. (2014, November). Stress, fatigue and automation: How do they intersect in UAV operation? *DoD HFE Virtual UxS SubTAG Meeting*.

Matthews, G., Neubauer, C., Saxby, D.J., & Wohleber, R.W. (in press). Fatigue and stress in the automated vehicle: Strategies for maintaining safety. *International Journal of Safety across High-Consequence Industries.*

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., Teo, G., Wohleber, R. W., & Lin, J. (2016) Resilient autonomous systems: Challenges and solutions. *Resilience Week 2016: Transforming the resilience of cognitive, cyber-physical systems*. Chicago, IL, August.

Matthews, G., Reinerman-Jones, L., Lin, J., Wohleber, R., Chiu, P., Calhoun, G., & Ruff, H. (2015). Individual differences in performance, trust, and stress during multi-UAV operator. *DoD HFE TAG Meeting 69*, Orlando, FL, May 2015.

Matthews, G., Reinerman-Jones, L., Wohleber, R., Lin, J., Mercado, J., & Abich, J. (2015). Workload is multidimensional, not unitary: What now? In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition* (pp. 44-55). New York: Springer.

Matthews, G., Wohleber, R., Lin, J., Calhoun, G., & Funke, G. (2016, May). Tracking fatigue and reliance on automation in Multi-UAV operation. *DoD HFE TAG Meeting 70*, Hampton, VA.

Russell, S. M., Funke, G. J., Miller, B. T., Dukes, A., Flach, J. M., Watamaniuk, S. N. J., Strang, A. J., Menke, L., & Brown, R. (2014). Alternative indices of performance: An exploration of eye gaze metrics in a visual puzzle task (Technical Report AFRL-RH-WP-TR-2014-0095). Air Force Research Laboratory, 711 Human Performance Wing, Human Effectiveness Directorate, Wright-Patterson AFB.

Tolston, M., Funke, G., Matthews, G., Wohleber, R., Lin, J., Calhoun, G., & Ruff, H. (2016, July). Effects of automation reliability on structure of gaze patterns over time. Poster to be presented at the *7th International Conference on Applied Human Factors and Ergonomics*, Orlando, FL.

Wohleber, R. W. (2016, May). *The impact of automation reliability and fatigue on reliance* (Doctoral dissertation). University of Central Florida. Retrieved from https://www.researchgate.net/publication/303518932_The_impact_of_automation_reliability_and_fatigue_on_reliance

Wohleber, R. W., Calhoun, G. L., Funke, G. J., Ruff, H. A., Chiu, C.-Y. P., Lin, J., & Matthews, G. (2016). The impact of automation reliability on performance and reliance changes with operator fatigue. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*.

Wohleber, R. W., Matthews, G., Funke, G. J., & Lin, J. (2016). Considerations in physiological metric selection for online detection of operator state: A case study. In *Proceedings of International Conference on Human-Computer Interaction.* Toronto, Canada: Springer-Verlag.

11. LIST OF ACRONYMS AND ABBREVIATIONS

Acronym	Definition
AFOSR	Air Force Office of Scientific Research
AFRL	Air Force Research Laboratory
ALOA	Adaptive Levels of Autonomy (referring to multi-RPA simulation used)
AOI	Area of Interest
DSSQ	Dundee State Stress Questionnaire
HCT	Human-Computer Trust
LOA	Level of Automation
PAC	Perceived Attentional Control
PERCLOS	Percent time eye closed
ROE	Rules of Engagement
RPA	Remotely Piloted Aircraft
RQA	Recurrence Quantification Analysis
s	Second(s)